Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Milutin Spasić

Algoritem za detekcijo komponent kompleksov proteinov v interakciji z RNA

diplomsko delo

univerzitetni študijski program prve stopnje
računalništvo in informatika

Mentor: doc. dr. Tomaž Curk
Somentor: prof. dr. Jernej Ule

Ljubljana 2014

Besedilo je oblikovano z urejevalnikom besedil LaTeX.

Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Tematika naloge:

Kompleksi proteinov in RNA imajo pomembno vlogo pri uravnavanju izražanja genov. V diplomski nalogi razvijte in implementirajte metodo, s katero bo možno odkriti vzorce vezave posameznih proteinov, ki sestavljajo kompleks z RNA. Vhod v metodo naj bodo podatki o mestih interakcije kompleksa ter vseh posameznih proteinov, za katere obstajajo eksperimentalni podatki. Izhod naj bodo vzorci, ki opisujejo kombinacije mest vezav posameznih proteinov, s katerimi najboljše opišemo mesta interakcije kompleksa. Metodo ovrednotite na podatkih o kompleksu *SmB*, za katerega obstaja predznanje o posameznih proteinskih komponentah.

Ribonucleoprotein complexes of proteins and RNA play an important role in gene expression regulation. In the thesis develop and implement a method, which will be able to identify the various protein components of a given complex. The method should accept protein-RNA interaction data on a given complex and all available protein-RNA interaction data on individual proteins. The method should output the combinations of individual proteins that best describe the interaction profile of the protein complex. Evaluate the method on the *SmB* complex, for which a number of components have been identified experimentally.

Izjava o avtorstvu diplomskega dela

Spodaj podpisani Milutin Spasić, z vpisno številko **63110397**, sem avtor diplomskega dela z naslovom:

Algoritem za detekcijo komponent kompleksov proteinov v interakciji z RNA

S svojim podpisom zagotavljam, da:

- sem diplomsko delo izdelal samostojno pod mentorstvom doc. dr. Tomaža Curka in somentorstvom prof. dr. Jerneja Uleta,
- so elektronska oblika diplomskega dela, naslov (slov., angl.), povzetek (slov., angl.) ter ključne besede (slov., angl.) identični s tiskano obliko diplomskega dela,
- soglašam z javno objavo elektronske oblike diplomskega dela na svetovnem spletu preko univerzitetnega spletnega arhiva.

V Ljubljani, dne 12. septembra 2014 Podpis avtorja:

Zahvaljujem se družini za vso ljubezen in podporo. Hvala najboljšim prijateljem za vso pomoč in podporo v času študija.

Posebna zahvala mentorju, doc. dr. Tomažu Curku, za vse znanje, ki ga je velikodušno delil z mano, in za vso pomoč pri izdelavi diplomskega dela.

Hvala tudi somentorju, prof. dr. Jerneju Uletu, za nasvete in pomoč pri izdelavi diplomskega dela.

Babici, ki je bila in oče in mati, ko je to bilo potrebno.

Kazalo

Seznam uporabljenih kratic

kratica	angleško	slovensko
iCLIP	individual nucleotide resolution UV crosslinking and immunoprecipitation	
NMF	non-negative matrix factorization	faktorizacija nenegativnih matrik
RNA	ribonucleic acid	ribonukleinska kislina
mRNA	messenger RNA	informacijska RNA
pre-mRNA	precursor mRNA	prekurzorska mRNA
DNA	deoxyribonucleic acid	deoksiribonukleinska kislina
cDNA	complementary DNA	komplement DNA

Povzetek

Proteini so pomembni akterji v mnogih celičnih procesih. Interakcije proteinov in RNA imajo pomembno vlogo pri uravnavanju izražanja genov in posledično njihove funkcije. V interakcijo z RNA navadno vstopa več tako posameznih proteinov kakor tudi skupin proteinov. Metoda iCLIP omogoča do nukleotida natančno detekcijo mest interakcij na RNA z izbranim proteinom. V diplomskem delu smo razvili metodo, ki sprejme nabor mest na RNA, ki so v interakciji z izbranim proteinskim kompleksom. V okolici teh mest nato poišče vzorce interakcij iz nabora interakcij proteinov z RNA, za katere imamo podatke. Najdeni vzorci vključujejo tako posamezne komponente (proteine) danega kompleksa kakor tudi proteine, ki danega kompleksa ne tvorijo, ampak vseeno vplivajo na interakcije kompleksa z RNA. Razvita metoda temelji na postopeku faktorizacije nenegativnih matrik. Metodo smo uspešno preizkusili na podatkih o kompleksu *SmB*, kjer smo uspešno določiti nekaj proteinov, ki so soudeleženi v kompleksu *SmB*.

Ključne besede: interakcije protein-RNA, faktorizacija nenegativnih matrik, iCLIP, kompleksi proteinov in RNA.

Abstract

Proteins play an important role in many processes in a cell. Protein-RNA interactions greatly affect the balancing of gene expressions and consequently their functions. Interaction with RNA may occur with a single protein or a protein complex. iCLIP method is able to detect the protein-RNA crosslink spots with nucleotide resolution. We developed a method that takes a set of crosslink spots that intersect with the choosen protein complex as an input. Method is searching the crosslink spots neighbourhood for the protein-RNA interaction patterns. Found patterns include parts (proteins) of the choosen complex and other proteins that are not a part of the choosen complex, but still affect the interactions of the choosen complex with the RNA. Method that we developed is based on the non-negative matrix factorization. We successfully tested the method on a *SmB* complex where we found a few proteins that cooporate with *SmB* complex.

Keywords: protein-RNA interactions, nonnegative matrix factorization, iCLIP, protein complexes.

Poglavje 1

Uvod

V zadnjih desetletjih je molekularna biologija začela izjemno hiter razvoj. Pojavila se je v 30. letih 20. stoletja kot povezava med do tokrat popolnoma ločenima vejama biologije: genetiko in biokemijo. Molekularni biologi so si zadali opisati strukturo, funkcijo in povezave dveh makromolekul: nukleinskih kislin in proteinov. Proteini se povezujejo z drugimi molekulami in tako ustvarjajo različne strukture ter tako regulirajo vse procese v organizmu. Zaradi tega so postali zelo zanimivi za raziskovanje. Za razumevanje celičnih procesov je najprej treba razumeti, kateri proteinski kompleksi so povezani s procesom, posamezne proteinske komponente, pa tudi ostale proteine, ki uravnavajo kompleks.

Kompleks *SmB* je eden izmed zelo pomembnih proteinskih kompleksov [4]. Pripada skupini sedmih RNA-vezavnih proteinov kompleksa *Sm* (Slika 1.1). Proteini komplesa *Sm* pomagajo pri ustvarjanju spajalnega telesca (angl. *spliceosome*). Spajalno telesce se ustvarja okrog pre-mRNA (prekurzorjev informacijske RNA, pre-mRNA), kjer izvršuje izrezovanje (angl. *splicing*) intronov in spajanje eksonov. Končni produkt tega procesa je mRNA (informacijska RNA). Kompleks *SmB* je prav tako sestavljen iz različnih proteinov, ki prihajajo v stik z RNA. Različni deli kompleksa *SmB* imajo različne biokemijske preference do vezave z določenimi krajšimi sekvencami RNA. Zaradi preference do določene sekvence nukleotidov in vseh ostalih molekul, ki so v interakciji z RNA v bližini te sekvence, se določeni proteini iz kompleksa *SmB* začnejo sopojavljati v podobnih regijah na RNA. Tako proteina *U2AF65* in *U2AF35* preferirata 3'-konec področja na stiku med intronom in eksonom. Proteina *TIA1* in *TIAL1* pa 5'-konec področja na stiku med eksonom in intronom. Raziskave so do sedaj pokazale, da sta *TIA* in *U2AF65* proteini del kompleksa *SmB* [9]. Ista študija je pokazala obstoj neznanih proteinov, ki tudi sestavljajo ta kompleks [9].

Slika 1.1: Proteinski kompleks *Sm* v interakciji z RNA. Prikazana je struktura 2Y9C iz zbirke PDB [8].

Cilja diplomske naloge sta dva:

1. Razviti metodo za dekompozicijo podatkov iCLIP o mestih vezave danega proteinskega kompleksa in RNA na vzorce, ki vključujejo podatke iCLIP o mestih vezave vrste proteinov. Najdeni vzorci naj bi vključevali tako neznane proteinske komponente danega kompleksa kakor tudi proteine, ki ne tvorijo danega kompleksa, ampak imajo vseeno vpliv na njegove interakcije z RNA.

2. Uporabiti metodo na podatkih iCLIP o kompleksu *SmB* in jo ovrednotiti s pomočjo predznanja o kompleksu *SmB*. Pri vrednotenju želimo preveriti, ali razvita metoda najde vse že znane komponente kompleksa *SmB*. Za vzorce, ki ne opisujejo znanih komponent (proteinov) kompleksa *SmB*, želimo preveriti, ali zanje obstaja smiselna biološka razlaga.

Poglavje 2

Interakcije med proteini in RNA

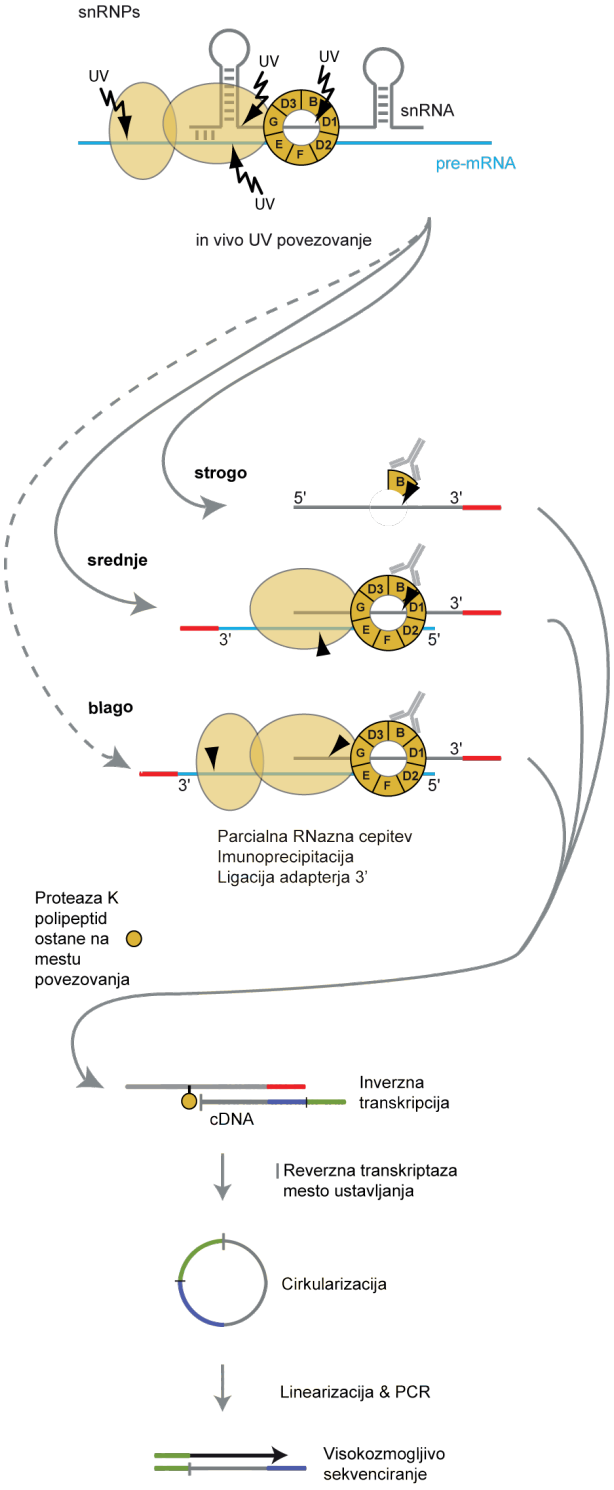
Podatkov o interakciji kompleksov proteinov in RNA je relativno malo. V zadnjem desetletju so razvili mnogo metod za detekcijo interakcij med proteini in RNA. Vsekakor novejše metode dosegajo večje natančnosti in na ta način zagotavljajo bolj zanesljive podatke. Pridobitev podatkov je še vedno relativno drag proces, kar je tudi razlog, zakaj je podatkov malo. V našem delu smo uporabili podatke iCLIP, ki so bili pridobljeni v laboratoriju prof. dr. Jerneja Uleta iz University College London (UCL).

Podatki so razdeljeni na dve osnovni skupini: podatki o kompleksu *SmB* in podatki o posameznih proteinih. Podatki o kompleksu *SmB* združujejo vse zaznane interakcije kompleksa *SmB* z RNA skozi različne faze celičnega cikla. Ostali proteini obsegajo 56 poskusov, ki opisujejo interakcije z RNA za 39 različnih proteinov ali njihovih mutacij. Vsa mesta iz ostalih proteinov so detektirana na istih področjih, kjer je detektiran tudi kompleks *SmB*. Uporabili smo datoteke v formatu bedGraph, s katerim so zabeležena mesta vezave posameznih proteinov in kompleksa *SmB*.

2.1 Metoda za detekcijo interakcij iCLIP

Metoda iCLIP se uporablja za določanje mest na RNA, ki vstopajo v interakcijo z izbranim proteinom. Z ultravijolično svetlobo C (UVC) osvetlijo žive celice in tako inducirajo ustvarjanje kovalentnih vezi (povezovanje UV, angl. *cross-linking*) na mestih, kjer protein prihaja v stik z RNA. Tako dobljen kompleks protein-RNA imunoprecipitirajo s pomočjo protitelesa (angl. *antibody*), ki je specifično za izbrani protein. Vezane RNA odstranijo iz kompleksov protein-RNA in jih nato sekvencirajo. Odčitke kartirajo na referenčni genom in tako določijo do nukleotida natančno mesta interakcij med izbranim proteinom ter RNA (Slika 2.1).

Pri izvajanju poskusa iCLIP je možno uporabiti različne koncentracije proteaz, s katerimi vplivamo na stopnjo fragmentiranja proteina (blaga, srednja, stroga), oziroma proteinskega kompleksa v interakciji z RNA.

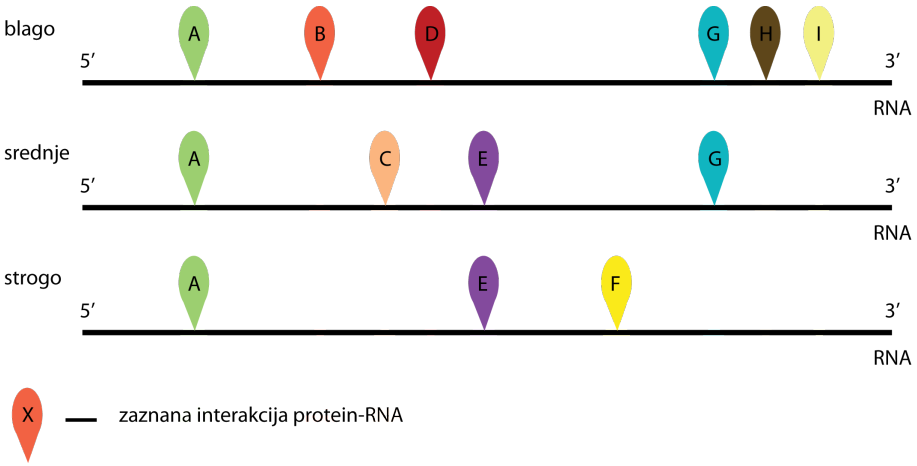


Slika 2.1: Metoda iCLIP za detekcijo mest na RNA v interakciji s proteinom.

Poleg mesta interakcije lahko v poskusih iCLIP določimo tudi moč oziroma po-

gostost interakcije proteina in posameznih mest na RNA. To vrednost imenujemo vrednost *cDNA*.

Odvisno od izbrane tehnike bo končni nabor vseboval manj ali več zaznanih mest interakcije. Bolj pomemben postopek je sledenje istim detektiranim mestom vezave za različne stopnje fragmentiranja. Tako lahko ugotovimo zanesljivost podatkov o posameznih zaznanih mestih. Slika 2.2 prikazuje zaznane interakcije protein-RNA, kjer ena črka oziroma barva predstavlja eno interakcijo. Mesta **B**, **C**, **D**, n **F**, **H** in **I** predstavljajo slabe interakcije ali celo šum v podatkih (**C** in **F**). Mesta interakcije, ki se obdržijo skozi več nivojev, so bolj zanesljiva. Tako je mesto interakcije **A** najbolj zanesljivo, mesti **E** in **G** pa sta nekoliko manj zanesljivi.



Slika 2.2: Zaznane interakcije glede na stopnjo fragmentiranja.

2.2 Format bedGraph

Format bedGraph je standardiziran [10] in se v bioinformatiki uporablja za shranjevanje podatkov o anotaciji posameznih mest (en nukleotid) ali intervalov (zaporedje nukleotidov) na genomu.

Vsaka vrstica v tekstovni datoteki formata bedGraph opisuje eno mesto oz. interval na genomu. Sestavljena je iz štirih stolpcev (ločenih s posebnim znakom *TAB*):

- kromosom,
- koordinata prvega nukleotida v intervalu,
- koordinata prvega nukleotida za zadnjim v intervalu in

- številska vrednost, ki je pripisana intervalu. Predznak navadno pove, za katero stran (verigo) DNA se zapis nanaša. V poskusih iCLIP je to število detektiranih interakcij oziroma vrednost *cDNA*.

2.3 Detekcija kompleksov več proteinov

Metoda iCLIP predstavlja revolucijo glede natančnosti in kakovosti mapiranja interakcij med proteinom in RNA. Kljub temu pa še vedno ni idealna. Učinkovitost povezovanja s svetlobo UV in posledično detekcije interakcij pri metodi iCLIP je manj kot 1%, kar pomeni, da z metodo ne zaznamo več kot 99% vseh interakcij. Slaba učinkovitost ima tudi druge posledice. Ker vsako interakcijo zaznamo z manj kot 1% verjetnostjo, je tudi verjetnost zaznavanja celotnih vzorcev proteinskih kompleksov zelo majhna. Drugače razloženo, tudi če metoda zazna vse interakcije izbranega proteina, je zelo majhna verjetnost, da bodo vse interakcije zaznane na isti RNA. Zaradi tega podatki prikazujejo večinoma delne vzorce vezave, iz katerih z osnovnimi metodami ni mogoče odkriti celotnega vzorca, temveč le posamezne dele. Za povezovanje teh delov vzorcev je treba uporabiti vsaj še informacijo o sekvenci nukleotidov, na katere se veže izbrani protein.

Drugi problem pri podatkih predstavlja gostota sekvenciranj pri metodi iCLIP. Podatki iCLIP o drugih proteinih so pridobljeni pod različnimi pogoji oz. parametri metode iCLIP. Zaradi tega so eni podatki bolj gosto sekvencirani kot drugi in vsebujejo več zaznanih interakcij (do 20-krat več). To povečuje možnost, da jih zaznamo v bližini nukleotidov, na katerih je zaznan tudi kompleks *SmB*. Razlike v velikostih naborov podatkov imajo velik vpliv na iskanje skupnih vzorcev interakcije. Tabela 2.1 prikazuje število zaznanih interakcij za vsak protein. Navedeno je tudi predznanje o tem, ali je posamezni protein del kompleksa *SmB*.

protein	**vezavna mesta**	**del kompleksa *SmB***
A3F	779664	
A3G	1840471	
Btz	1348595	
CGI99	333356	
CPSF160	681629	
Celf2	436964	
Celf4	1076472	
Nadaljuje se na naslednji strani		

Tabela 2.1 – nadaljevanje prejšnje strani

protein	**vezavna mesta**	**del kompleksa *SmB***
FUS	391422	
Hexim1	108487	
HuR	5465572	
IGF2BP1	555435	
IGF2BP2	182590	
IGF2BP3	1120148	
Matr3	966798	
Musashi	16205	
NSUN2	25204	
Nova	69217	
PTB	12899108	
Pura	9570	
RBM7FLAG	272078	
RNPS1	1045762	
SAFB1	146123	
STAU1	707928	
Sam68	45373	
SmB	11708470	
TDP43	619087	
TIA1	1140676	ja
TIAL1	982558	ja
TRA2B	822785	
U2AF65	45035683	ja
Upf3b	690308	
bisox	538858	
eIF4E	485173	
hSRSF2	170957	
hSRSF6	118895	
hnRNPA1	13869999	
hnRNPC	5676781	
hnRNPK	62080	
hnRNPL	208547	
hnRNPLlike	26639	

Tabela 2.1: Število zaznanih interakcij posameznega proteina.

Poglavje 3

Metode

Algoritem 1, ki smo ga razvili je sestavljen iz več glavnih korakov:

1. osnovno predprocesiranje podatkov (glej prvo vrstico algoritma 1 in podpoglavje 3.1)
2. delitev na genomske regije (del predprocesiranja, glej tretjo vrstico algoritma 1 in podpoglavje 3.1.1)
3. grajenje matrike X (glej četrto vrstico algoritma 1 in podpoglavje 3.2.2)
4. iskanje ranga faktorizacije (glej peto vrstico algoritma 1 in podpoglavje 3.2.3)
5. faktoriziranje matrike X (glej šesto vrstico algoritma 1 in podpoglavje 3.2)
6. iskanje vzorcev interakcij med proteini in RNA (glej sedmo vrstico algoritma 1 in podpoglavje 3.2.2)
7. vizualizacija (glej osmo vrstico algoritma 1 in podpoglavje 3.5)

3.1 Predprocesiranje podatkov

Podatki o kompleksu *SmB*, ki jih uporabljamo za vrednotenje naše metode, so že dodatno procesirani z upoštevanjem biološkega predznanja. Celični cikel obsega 5 različnih faz. V vsaki od teh je bil izveden poskus iCLIP za detekcijo interakcij kompleksa *SmB* in RNA. Tako pridobljene podatke so biologi združili in ustvarili reprezentativno množico interakcij kompleksa *SmB* ter RNA za celoten celični cikel. Podatke so procesirali v laboratoriju prof. dr. Jerneja Uleta na UCL. Prejeli smo jih v obliki ene datoteke formata bedGraph, ki vsebuje približno 12 M vrstic, kjer vsaka vrstica predstavlja eno mesto interakcije kompleksa *SmB* na

RNA. Biološki podatki po navadi vsebujejo veliko šuma. Kot vse druge podatke je tudi te bilo treba dodatno procesirati in izbrati podmnožico najbolj informativnih. Drugi razlog za izbiro manjše podmnožice podatkov je časovna kompleksnost uporabljenih računskih metod.

Algorithm 1 Celoten program

Vhod:

data : vhodni podatki o mestih interakcije.

okvir : območje, znotraj katerega gledamo gostoto proteinov.

meja_vzorcenja : maksimalno število mest, ki jih naključno izberemo.

min_cDNA : minimalna vrednost *cDNA* za izbiro podatkov.

max_pos : maksimalno število mest, ki jih metoda opiše.

okvir_glajenja : območje, v katerem pozicijam pripišemo vrednost *cDNA* referenčne interakcije.

max_rang : maksimalna vrednost ranga faktorizacije.

top_n : maksimalno število najboljših vzorcev, ki jih prikažemo.

okvir_skupine_interakcij : območje, znotraj katerega vsa mesta interakcije pripišemo enemu predstavniku območja.

```
podatki_po_regijah = predprocesiranje(data, okvir,
    meja_vzocenja, min_cDNA, max_pos)
for regija in podatki_po_regijah:
    zglajeni_podatki = razdeli_in_zgladi(regija,
    okvir_glajenja)
    X = zgradi_matriko_X(zglajeni podatki, okvir)
    rang_faktorizacije = najdi_rang(X, max_rang)
    W, H = nmf(X, rang_faktorizacije)
    vsi_vzorci = najdi_vse_vzorce(X, W, H, top_n,
    okvir_skupine_interakcij)
   vizualizacija(vsi_vzorci)
```

3.1.1 Izbira mest interakcij

Izbira mest je prvi korak predprocesiranja podatkov, kjer se odstrani večina šuma in izberejo zanesljiva mesta interakcij.

Naključno vzorčenje mest interakcij

Razlika v velikosti naborov podatkov o posameznih proteinih pomembno vpliva na nadaljnjo izbiro pozicij. Poskusi, ki vsebujejo več zaznanih interakcij, imajo tudi večjo verjetnost, da so v bližini mest kompleksa *SmB*, ki ga želimo opisati. Ta problem smo rešili z vzorčenjem podatkov, kjer smo za vsak protein naključno izbrali podmnožico največ dveh milijonov mest interakcije proteina in RNA. Na ta način obdržimo vsa mesta interakcij v poskusih, kjer je detektiranih mest malo. Hkrati zmanjšamo podatke o poskusih, kot so *U2AF65*, *PTB* in *hnRNP* (glej tabelo 2.1). Na ta način smo izenačili možnosti vseh proteinov, da so njihova vezavna mesta v bližini mest interakcije ciljnega kompleksa *SmB*.

Vrednost *cDNA*

Vrednost *cDNA* predstavlja moč interakcije med proteinom in RNA. Mesta z višjo vrednostjo *cDNA* predstavljajo bolj zanesljiva mesta interakcije, saj je bila tam interakcija večkrat detektirana. Nespecifične interakcije, kjer se opazovani protein slučajno poveže z nekim mestom na RNA, imajo nizke vrednosti *cDNA*. Takšne interakcije so dokaj pogoste, ampak tudi enakomerno razpršene po celotnem genomu. Za določitev spodnje meje za vrednost *cDNA* smo preverili rezultate, ki jih dobimo za meje ena, dva in tri. Po pregledovanju rezultatov in posvetovanju z avtorji metode iCLIP smo se odločili obravnavati le vezavna mesta, kjer je vrednost *cDNA* vsaj dva (glej četrto vrstico algoritma 2). Za podrobnosti glej podpoglavje 4.1. Tabela 3.1 prikazuje število vseh vezavnih mest posameznega proteina in število mest s $cDNA \geq 2$.

protein	**poskus iCLIP**	**vsa vezavna mesta**	**vezavna mesta s *cDNA* ≥ 2**
A3F	A3F-GFP-CEMSS-hg19_4031	309094	60273
A3F	A3F-T7-CEMSS-hg19_3952	553109	113045
A3G	A3G-GFP-CEMSS-hg19_3942	1486603	407214
A3G	A3G-T7-CEMSS-hg19_3995	610453	136948
Btz	Btz-HeLa-hg19_3955	1348595	489541
CGI99	CGI99-FlpIn293-hg19_4033	333356	69436
CPSF160	CPSF160-Hela-4SU-hg19_4036	681629	173725
Celf2	Celf2-293Flp-hg19_3966	436964	95860

Nadaljuje se na naslednji strani.

Tabela 3.1 – nadaljevanje prejšnje strani

protein	**poskus iCLIP**	**vsa vezavna mesta**	**vezavna mesta s *cDNA* ≥ 2**
Celf4	Celf4-FlpIn293-hg19_4016	1076472	350280
FUS	FUS-ES-hg19_3945	194055	40569
FUS	FUS-GFP-Hela-hg19_3967	37916	20074
FUS	FUS-Hela-hg19_4005	12470	4638
FUS	FUS-SHSY5Y-hg19_4020	20069	6278
FUS	FUS-brain-hg19_3998	149129	32206
Hexim1	Hexim1-FLAG-FlpIn293-hg19_3972	108487	35652
HuR	HuR-FlpIn293-hg19_4361	5465572	2003446
IGF2BP1	IGF2BP1-Hela-hg19_3962	555435	123690
IGF2BP2	IGF2BP2-Hela-hg19_4006	182590	37926
IGF2BP3	IGF2BP3-Hela-hg19_3947	1120148	258888
Matr3	Matr3-Hela-hg19_4029	966798	67014
Musashi	Musashi-U251-hg19_4019	14896	4031
Musashi	Musashi-hESdiff-hg19_3943	1938	545
NSUN2	NSUN2-293-hg19_4007	25204	17457
Nova	Nova-Brain-hg19_3968	69217	13770
PTB	PTB-Hela-hg19_3987	12899108	5472608
Pura	Pura-Hela-hg19_3963	9570	3640
RBM7FLAG	RBM7FLAG-293-hg19_3960	272078	60626
RNPS1	RNPS1-Hela-hg19_4002	1045762	392695
SAFB1	SAFB1-MCF7-hg19_3994	106912	35470
SAFB1	SAFB1-MDAMB231-hg19_4012	44451	13362
STAU1	STAU1-FlpIn293-hg19_4023	707928	165510
Sam68	Sam68-MDAMB231-hg19_4008	45373	12689
SmB	SmB-all-ref	11708470	4208154
TDP43	TDP43-ES-hg19_3973	527662	141139
TDP43	TDP43-Hela-hg19_3954	7063	1807
TDP43	TDP43-SHSY5Y-hg19_3979	127329	31245
TIA1	TIA1-Hela-hg19_3959	1140676	371882
TIAL1	TIAL1-Hela-hg19_3936	982558	314506
TRA2B	TRA2B-MCF7-hg19_3993	158730	34853
TRA2B	TRA2B-MDAMB231-hg19_3958	715676	183212

Nadaljuje se na naslednji strani.

Tabela 3.1 – nadaljevanje prejšnje strani

protein	**poskus iCLIP**	**vsa vezavna mesta**	**vezavna mesta s *cDNA* ≥ 2**
U2AF65	U2AF65-FlpIn293-hg19_4015	154452	44962
U2AF65	U2AF65-Hela-hnRNPCkd-hg19_3999	19483467	9048424
U2AF65	U2AF65-Hela-wt1-hg19_4014	15539106	7594931
U2AF65	U2AF65-Hela-wt2-hg19_3949	28436319	18568726
Upf3b	Upf3b-Hela-hg19_4010	690308	229172
bisox	bisox-293-hg19_4001	538858	111014
eIF4E	eIF4E-FlpIn293-4SU-hg19_3997	235365	62370
eIF4E	eIF4E-FlpIn293-hg19_4030	302638	70110
hSRSF2	hSRSF2-Hela-hg19_4038	170957	43677
hSRSF6	hSRSF6-Hela-hg19_3980	118895	30722
hnRNPA1	hnRNPA1-Hela-hg19_4037	13869999	9052806
hnRNPC	hnRNPC-Hela-hg19_3981	5676781	1772633
hnRNPK	hnRNPK-FlpIn293-hg19_3935	62080	18577
hnRNPL	hnRNPL-Hela-hg19_3975	197388	54943
hnRNPL	hnRNPL-U266-hg19_3986	18482	7022
hnRNPLlike	hnRNPLlike-U266-hg19_4000	26639	9413

Tabela 3.1: Število vseh detektiranih vezavnih mest in mest s $cDNA \geq 2$ v posameznem poskusu iCLIP.

Tipi genomskih regij

Proteini vstopajo v interakcijo s celotnim pre-mRNA. Na določenih področjih so bolj prisotni kot na drugih. Eni proteini se tako pojavljajo strogo v eksonskih področjih, drugi samo globoko v intronih in daleč od eksona, tretji pa povsod. Zaradi te lastnosti proteinov smo se odločili, da podatke o mestih interakcij razdelimo na štiri skupine, in sicer glede na to, v kateri genomski regiji se pojavljajo (glej deveto vrstico algoritma 2):

- globoko v eksonu (ge),
- globoko v intronu (gi),
- na meji ekson-intron (ei),

- na meji intron-ekson (ie).

S takšno delitvijo dobimo skupine, ki jih je lažje razumeti. Na primer: za protein, ki se z visoko frekvenco pojavlja na vseh področjih, potem ne moremo trditi, da je pomembno povezan z drugimi proteini, ki se pojavljajo le na določenih področjih.

Druga prednost delitve analize na tip genomske regije je, da lahko za posamezne proteine, ki so del kompleksa *SmB*, določimo genomske regije, kjer se pogosto pojavijo. Vzorce pojavitve lahko nato primerjamo z odkritimi vzorci v ostalih, nespecifičnih regijah (glej podpoglavje 4.3). Tabela 3.2 prikazuje delež zaznanih interakcij obravnavanih proteinov v posameznih genomskih regijah in skupno število mest interakcije, ki imajo vrednost *cDNA* vsaj dva.

protein	**vezavna mesta s *cDNA* ≥ 2**	**% ei**	**% ie**	**% gi**	**% ge**
A3F	152013	3.45	17.34	18.62	60.59
A3G	474778	1.56	21.48	8.58	68.37
Btz	489541	3.43	28.09	2.63	65.85
CGI99	69436	1.03	32.28	5.44	61.25
CPSF160	173725	13.19	9.53	49.25	28.03
Celf2	95860	7.36	9.10	45.32	38.21
Celf4	350280	2.36	3.87	18.49	75.28
FUS	93423	3.28	8.99	28.69	59.03
Hexim1	35652	7.35	10.97	48.08	33.60
HuR	2003446	8.39	8.29	67.38	15.94
IGF2BP1	123690	3.79	9.11	31.38	55.72
IGF2BP2	37926	0.93	9.80	11.99	77.29
IGF2BP3	258888	2.56	11.03	24.40	62.01
Matr3	67014	1.79	2.79	85.74	9.68
Musashi	4332	2.53	11.39	44.30	41.77
NSUN2	17457	2.19	26.64	51.29	19.88
Nova	13770	0.00	1.54	5.82	92.64
PTB	5472608	8.44	7.00	64.23	20.33
Pura	3640	3.00	5.00	64.00	28.00
RBM7FLAG	60626	5.80	7.09	52.46	34.65
RNPS1	392695	5.63	24.07	6.84	63.46
SAFB1	46802	0.92	15.04	12.91	71.13

Nadaljuje se na naslednji strani.

Tabela 3.2 – Nadaljevanje prejšnje strani.

protein	vezavna mesta s *cDNA* ≥ 2	% ei	% ie	% gi	% ge
STAU1	165510	1.89	4.72	66.59	26.80
Sam68	12689	0.94	4.16	6.91	87.99
SmB	4208154	18.74	26.34	27.20	27.72
TDP43	163839	7.73	8.61	70.01	13.65
TIA1	371882	32.15	6.26	23.15	38.45
TIAL1	314506	36.57	7.00	20.42	36.01
TRA2B	204251	3.65	31.94	11.65	52.76
U2AF65	26033595	13.54	32.32	30.98	23.17
Upf3b	229172	6.66	25.93	6.11	61.29
bisox	111014	1.84	19.51	9.77	68.88
eIF4E	115925	3.63	23.51	6.08	66.78
hSRSF2	43677	8.21	30.23	13.10	48.45
hSRSF6	30722	4.84	28.28	11.31	55.57
hnRNPA1	9052806	7.01	8.97	53.36	30.66
hnRNPC	1772633	5.24	7.35	64.16	23.26
hnRNPK	18577	5.47	5.80	36.11	52.63
hnRNPL	58176	1.14	12.04	9.34	77.48
hnRNPLlike	9413	0.30	13.15	10.71	75.85

Tabela 3.2: Mesta interakcije s *cDNA* ≥ 2 in delež mest v genomskih regijah.

Eksoni in introni

Eksoni so zaporedja nukleotidov znotraj gena, ki določajo zaporedje mRNA [5]. Zaporedje v eksonih določa aminokislinsko sestavo tvorjenih proteinov. Združevanje eksonov poteka v postopku izrezovanja intronov, ki so nekodirajoča zaporedja nukleotidov in ločujejo eksone [5]. Geni, ki so zgrajeni iz eksonov in intronov, so značilni le za evkarionte.

Algorithm 2 predprocesiranje()

Vhod:

data : vhodni podatki o mestih interakcije.

okvir : območje, znotraj katerega gledamo gostoto proteinov.

meja_vzorcenja : maksimalno število mest, ki jih naključno izberemo.

min_cDNA : minimalna vrednost *cDNA* za izbiro podatkov.

max_pos : maksimalno število mest, ki jih metoda opiše.

Opis:

Metoda, ki naredi predprocesiranje podatkov (za klic metode glej prvo vrstico algoritma 1).

```
poskusi = preberi_poskuse()
poskusi = vzorcenje(poskusi, meja_vzorcenja, okvir)
for pozicija in data:
  if vrni_cDNA(pozicija) > min_cDNA - 1:
    izbrane_pozicije.add(pozicija)
  end
end

pozicije_po_regijah = razdeli_pozicije_po_regijah(
   izbrane_pozicije)
for regija in pozicije_po_regijah:
  regija = regija.sort()[:max_pos]
end
return pozicije_po_regijah
```

3.1.2 Glajenje podatkov

Čeprav metoda iCLIP omogoča pridobivanje popolnoma zanesljivih podatkov z nukleotidno natančnostjo o mestih interakcije proteina in RNA, biološki procesi interakcij protein-RNA mnogokrat ne potekajo z nukleotidno natančnostjo. Ker lahko tudi sosednja mesta na RNA vstopajo v interakcijo s proteinom, smo podatke o mestih vezave zgladili in tako omogočili odstopanja za **n**-nukleotidov. Vrednost parametra **n** smo postavili na dva, kar pomeni, da vsem nukleotidom, ki so največ dve poziciji stran od mesta referenčne interakcije, pripišemo vrednost *cDNA* referenčne interakcije. Tako lahko opišemo veliko več mest. Vrednost parametra **n** smo določili na podlagi predznanja in izkušenj avtorjev metode iCLIP.

Algoritem 3 prikazuje postopek delitve podatkov in glajenje. Podatki so bili razdeljeni glede na stran verige (angl. *strand*) in kromosom, ki mu pripadajo. Razdeljene podatke smo zgladili tako, da smo vsem nukleotidom znotraj okvirja velikosti dva pripisali podatek o moči interakcije (*cDNA*).

Algorithm 3 razdeli_in_zgladi_podatke()

Vhod:

data : originalni vhodni podatki.

okvir_glajenja : okvir, znotraj katerega zgladimo podatke.

Opis:

Metoda, ki podatke razdeli glede na trak in kromosom ter potem zgladi (za klic metode glej tretjo vrstico algoritma 1).

```
for pozicija in data:
  trak = vrni_trak(pozicija)
  kromosom = vrni_kromosom(pozicija)
  zglajene_pozicije = zgladi_pozicijo(pozicija,
   okvir_glajenja)
  slovar_nukleotidov[trak][kromosom].add(zglajene_pozicije)
end
return slovar_nukleotidov
```

3.2 Faktorizacija nenegativnih matrik (NMF)

Faktorizaija nenegativnih matrik (NMF) je matematična, optimizacijska metoda [3, 7], ki vhodno matriko $\mathbf{X}$ faktorizira v dve manjši matriki $\mathbf{W}$ in $\mathbf{H}$, kjer velja:

$$\mathbf{X} = \mathbf{WH} \tag{3.1}$$

$$\mathbf{X}, \mathbf{W}, \mathbf{H} \geq 0,$$

kjer je $\mathbf{X} \in \mathbb{R}^{m \times n}, \mathbf{W} \in \mathbb{R}^{m \times k}, \mathbf{H} \in \mathbb{R}^{k \times n}, \mathbf{k} \ll m, n$

Problem faktorizacije matrike v splošnem ni rešljiv, ampak je predstavljen kot numerična aproksimacija. Pri matriki $\mathbf{X}$ vrstice predstavljajo vzorce, stolpci pa atribute. Zahtevana nenegativnost matrik nam omogoča lažjo interpretacijo

faktorjev v matrikah **W** in **H**. Tako je vsak vzorec v matriki **X** možno zapisati kot nenegativno linearno kombinacijo faktorjevrjev (3.2).

$$\mathbf{x}_{i,j} = \sum_{z=0}^{k} \mathbf{W}_{i,z}\mathbf{H}_{z,j} \tag{3.2}$$

Gručenje

NMF ima lastnost, da med faktorizacijo avtomatično naredi tudi gručenje vzorcev v matriki **W**. Vzorci, ki imajo veliko skupnega, imajo podobne profile v matriki **W**. Z druge strani nam matrika **H** s stališča atributov razkrije pomembne značilnosti posameznih gruč. Aproksimacijo matrike **X** dosežemo z minimizacijo funkcije napake **J** (3.3).

$$\mathbf{J} = \min_{\mathbf{W},\mathbf{H}} \|\mathbf{X} - \mathbf{WH}\|_{\mathbf{F}} \tag{3.3}$$
$$\mathbf{X}, \mathbf{W}, \mathbf{H} \geq 0$$

Gručenje pri NMF ni povsem samoumevno, vendar če bi zgornji funkciji dodali še omejitev $\mathbf{HH}^T = I$, bi dobili postopek minimizacije algoritma ***k-means clustering***. Zaradi zahtevane nenegativnosti lahko direktno preslikamo vrednost faktorja v moč pripadanja posameznega vzorca določeni skupini matrike **W**. Tako najvišja vrednost faktorja v stolpcu **i** matrike **H** opredeli, katero gručo opisuje stolpec **i**.

3.2.1 Uporaba testa Z za izbiro celic v matrikah W in H

Z-test (3.4), oziroma Z-vrednost, je statistična mera, ki prikaže razmerje med povprečjem skupine in posameznim elementom skupine [5]. Z-vrednost podaja, za koliko standardnih odklonov neka vrednost odstopa od povprečja. Element z Z-vrednostjo nič ima vrednost povprečja. Z-vrednost je lahko pozitivna ali negativna, odvisno od tega, ali je vrednost elementa manjša ali večja od povprečja.

$$z = \frac{\mathbf{x} - \mu}{\sigma}, \tag{3.4}$$

kjer je $\mathbf{x}$ *trenutna vrednost,* μ *povprečna vrednost (3.5)*
in σ *standardni odklon (3.6).*

$$\mu = \frac{\mathbf{1}}{\mathbf{N}} \sum_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{x_i}, \tag{3.5}$$

kjer je $\mathbf{x_i}$ *i-ti primer in* $\mathbf{N}$ *število vseh primerov.*

$$\sigma = \sqrt{\frac{\mathbf{1}}{\mathbf{N}} \sum_{\mathbf{i=1}}^{\mathbf{N}} (\mathbf{x_i} - \mu)}, \tag{3.6}$$

kjer je $\mathbf{x_i}$ *i-ti primer,* μ *povprečna vrednost in* $\mathbf{N}$ *število vseh primerov.*

Pri NMF smo uporabili Z-vrednost za ocenjevanje pripadnosti določenega atributa in pozicije posameznim skupinam. Takšen postopek se imenuje rangiranje. Pri rangiranju vse vrednosti v stolpcih matrike **H** in vrsticah matrike **W** označimo s celimi številkami od ena do **k**, in sicer od najnižje Z-vrednosti do najvišje. Potem za vsako skupino, torej vrstico v matriki **H**, poiščemo elemente z najvišjim rangom, **k**, in te atribute izberemo kot tiste, ki najbolje opisujejo opazovano skupino. Pri pozicijah v matriki **W** poiščemo skupino z najvišjim rangom in ji pripišemo to pozicijo. Takšen način ocenjevanja pripadnosti je mogoč zaradi vpeljane nenegativnosti pri NMF, kjer vrednost vsake številke pomeni, koliko ta prispeva h končni vrednosti posamezne celice v **X** matriki.

3.2.2 Uporaba NMF za detekcijo proteinskih kompleksov

NMF se je že pokazala kot zelo primerna za reševanje problemov v bioinformatiki, in sicer predvsem zaradi možnosti relativno enostavnega dodajanja novih podatkovnih virov v napovedni model. V matriki, ki smo jo zgradili, smo za atribute postavili posamezne proteine, vrstice pa so predstavljale posamezne nukleotide, kjer je zaznana interakcija referenčnega poskusa. V tej matriki smo z NMF iskali skupine podobnih vrstic, kjer se določeni proteini sopojavljajo.

V matriki **X** smo interakcije predstavili binarno. Če na razdalji **j** od mesta interakcije **i** zaznamo interakcijo s proteinom, potem v celico $\mathbf{x_{i,j}}$ vpišemo ena. V nasprotnem primeru pa nič. Binarizacija podatkov se je pokazala kot boljša metoda od obravnavanja vrednosti *cDNA*, ker so vrednosti *cDNA* med različnimi geni neprimerljive. V genih, ki so bolj izraženi, bomo detektirali več interakcij. To pa ne pomeni, da se opazovani protein raje veže na mRNA omenjenih genov.

V matriki **X** smo potem odstranili ničelne stolpce, ker ne prinašajo dodatnih informacij o interakcijah. S pomočjo pravilne izbire parametrov (glej podpoglavje 3.2.3) smo nad matriko **X** zagnali postopek NMF, ki nam je vrnil dekompozicijo v obliki matrik **W** in **H**. Nad matriko **W** smo zagnali postopek rangiranja po vrsticah s pomočjo Z-vrednosti (glej podpoglavje 3.2.1), ki vsako vrstico pripiše ustrezni gruči. Nad matriko **H** smo zagnali postopek rangiranja po

stolpcih s pomočjo Z-vrednosti (glej podpoglavje 3.2.1) in tako določili proteine, ki najbolje opisujejo vsako posamezno skupino.

Matrična faktorizacija nam poleg odkrivanja skupine poda še profil te skupine. Zelo verjetno, da pri kakšni izmed referenčnih pozicij eden od proteinov (atributov) manjka zaradi slabega sekvenciranja oziroma da se je kakšen protein naključno povezal in ne pripada vzorcu. Zaradi tega so določene vrstice v realnih podatkih imele vrednost ena pri atributih, ki se potem niso pojavljali v skupini in obratno. Vsem referenčnim mestom, ki pripadajo skupini $\mathbf{k_i}$, smo pripisali kombinacijo proteinov, ki opisuje skupino $\mathbf{k_i}$. Zaradi velikega števila atributov je tudi kombinacij teh atributov veliko. Zato smo pri vseh pozicijah obdržali samo tiste atribute, ki se pojavljajo v skupini, ki ji je dodeljena ta pozicija. Tako je vsaki vrstici v matriki $\mathbf{X}$ pripisana podmnožica atributov, ki jo opisujejo. Vse ponovitve teh kombinacij smo prešteli znotraj vsake skupine in izbrali pet najbolj pogostih kombinacij (glej algoritem 6). Tako pripravljene podatke smo podali metodi za prikaz rezultatov 3.5.

Algorithm 4 zgradi_matriko_X()

Vhod:

podatki : predprocesirani podatki.

okvir : območje, znotraj katerega gledamo gostoto proteinov.

Opis:

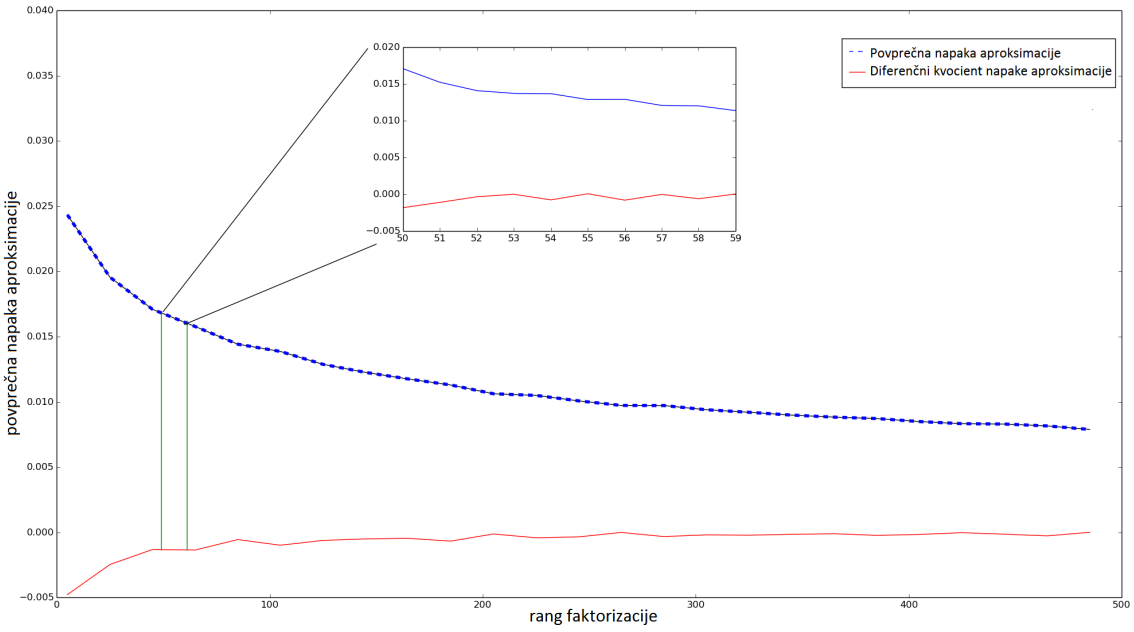
Metoda, ki zgradi matriko X, nad katero potem zaženemo postopek NMF (za klic metode glej četrto vristico algoritma 1).

```
poskusi = preberi_poskuse() # podatki o ostalih proteinskih
    kompleksih pridobijeni v poskusih iCLIP.
X = zeros(len(podatki), 81 * len(poskusi))

for i,podatek in enumerate(podatki):
  for j,poskus in enumerate(poskusi):
      presek = bisect(poskus, podatek.pozicija-okvir/2,
                      podatek.pozicija + okvir/2)
    X[i, presek.pozicije-podatek.pozicija + j*okvir] = 1

return X
```

3.2.3 Izbira ranga faktorizacije

Vrednost parametra **k** zelo vpliva na rezultate postopka NMF. Parameter **k** naj bi bil bistveno manjši od dimenzij (**m** in **n**) matrike **X** ter naj bi odražal število skupin v podatkih. Prevelika vrednost parametra **k** povzroči delitev na manjše skupine. Če uporabimo premajhno vrednost parametra **k**, pa dobimo manj čiste skupine, ki so posledično tudi manj informativne. Pravilno oz. pričakovano število skupin v večini primerov ni znano. Zato parameter **k** določimo na podlagi napake aproksimacije. S povečanjem parametra **k** se napaka aproksimacije zmanjšuje. Tipična oblika grafa napake aproksimacije je tako imenovana hokejska palica (angl. *hockey stick*). Na primeru, na sliki 3.1, napaka na začetku hitro pada. Po navadi za parameter **k** izberemo tisto vrednost, kjer se vrednost napake začne zmanjševati počasneje.

Zaradi stohastične komponente postopka NMF je iskanje ustreznega parametra **k** (glej algoritem 5) smiselno izvesti večkrat in nato izračunati povprečno napako aproksimacije. Na podlagi grafov tovrstnih povprečnih napak smo izbrali vrednosti za parameter **k**.



Slika 3.1: Povprečna napaka aproksimacije pri danem rangu faktorizacije.

Algorithm 5 najdi_rang()

Vhod:

X : matrika, ki jo želimo faktorizirati.

max_rang : maksimalna vrednost ranga faktorizacije, do katere bomo iskali najboljši rang.

Opis:

Metoda, ki poišče najboljši rang nenegativne matrične faktorizacije (za klic metode glej peto vrstico algoritma 1).

```
for i = 1 to max_rang do
  W, H = nmf(X, rang = i)
   aproksimacija_X = W * H
   napake_aproksimacije.add(mean(X-aproksimacija_X))
end
narisi_graf(napake_aproksimacije)

/* Uporabnik izbere mesto na grafu, kjer iz hitrega opadanja
preide v pocasno opadanje oziroma na grafu najde obliko
"hokejska palica" (angl. hockey stick). */

return izbrani_rang
```

Algorithm 6 najdi_vse_vzorce()

Vhod:

X : matrika, ki smo jo faktorizirali.

W : matrika, ki jo vrne nenegativna matrična faktorizacija.

H : matrika, ki jo vrne nenegativna matrična faktorizacija.

top_n : maksimalno število najboljših vzorcev, ki jih prikažemo.

okvir_skupine_interakcij : območje, znotraj katerega vsa mesta interakcije pripišemo enemu predstavniku območja.

Opis:

Metoda, ki najde najbolj frekventne vzorce proteinov za vsako posamezno skupino nukleotidov (za klic metode glej sedmo vrstico algoritma 1).

```
1 Wz = z_rangiranje(W)
2 Hz = z_rangiranje(H)
3 X = popravi_X(Wz, Hz, X)
4 for group in Hz:
5    pozicije = vrni_pozicije(W,X,group)
6    vzorci = najdi_vzorce(pozicije, okvir_skupine_interakcij)
7    vzorci.sort()[:top_n]
8    vsi_vzorci.add(vzorci)
9 return vsi_vzorci
```

3.3 Izčrpno iskanje

Izčrpno iskanje je ena izmed najbolj preprostih metod, ki jih lahko uporabimo pri reševanju danega problema. Zaradi preproste implementacije se uporablja kot referenčna metoda pri ocenjevanju uspešnosti drugih metod. Izčrpno iskanje preišče vse možnosti, kar je mnogokrat časovno zelo zahtevno. Kompleksnost izčrpnega iskanja je odvisna od dolžine vzorca, ki ga iščemo (3.7). V primerjavi z izčrpnim iskanjem je naša implementacija NMF 10-krat hitrejša.

$$T = \sum_{\mathbf{i}=1}^{\mathbf{n}} \sum_{\mathbf{j}=1}^{\mathbf{z}} \binom{\mathbf{f(i)}}{\mathbf{j}} \tag{3.7}$$

kjer je ***n*** *število primerov,* ***z*** *maksimalna dolžina kombinacije, ki jo iščemo, in* ***f(i)*** *število možnih elementov, ki pridejo v poštev za kombinacijo v vrstici* ***i****.*

Za faktorizacijo matrike s sto tisoč vrsticami in šest tisoč stolpci NMF potrebuje nekaj minut, da vrne vse odkrite vzorce. Vzorce nato sortiramo glede na frekvenco pojavitve vzorcev v podatkih. Izčrpno iskanje je pri isti matriki potrebovalo 30 minut. Rezultati izčrpnega iskanja so vse kombinacije dolžine od ena do maksimalno določene vrednosti. Kombinacije sortiramo glede na število vrstic (centrih kompleksov *SmB*), v katerih se pojavijo. Izčrpno iskanje poišče veliko več kombinacij kot NMF, in sicer predvsem zato, ker pri izčrpnem iskanju postopek popravljanja vhodne matrike **X** ni prisoten, kot to počnemo pri NMF s pomočjo matrike **H**. Drugi razlog za večje število kombinacij je način iskanja samih kombinacij. Pri izčrpnem iskanju preverimo vse kombinacije, ki jih lahko ustvarimo iz elementov ene vrstice, pri NMF pa celotno vrstico vzamemo kot eno samo kombinacijo.

3.4 Permutacijski test

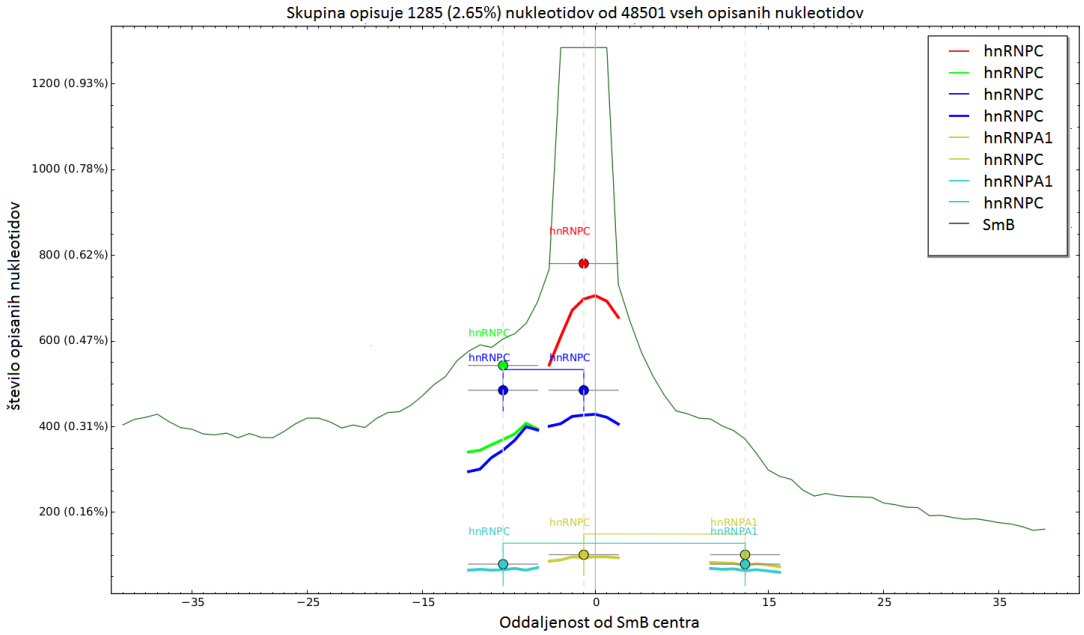
Za preverjanje značilnosti odkritih vzorcev z našo metodo smo uporabili permutacijski test. Celoten postopek smo ponovili tridesetkrat. Vsakič smo originalne podatke o kompleksu *SmB* naključno premaknili za ±200 nukleotidov. Na tako pridobljenih podatkih smo pognali naš algoritem, ki je poiskal vzorce in izračunal deleže vzorcev v posameznih genomskih regijah. Povprečne vrednosti tridesetih permutacij smo uporabili kot referenčne vrednosti, s katerimi smo primerjali rezultate, pridobljene na pravih podatkih. Rezultati permutacijskega testa so prikazani pod rezultati na pravih podatkih (glej poglavje 4) in predstavljajo povprečno vrednost vseh tridesetih naključnih permutacij.

3.5 Prikaz detektiranih proteinskih kompleksov

Naša metoda generira mnogo rezultatov, kar predstavlja izziv za interpretacijo. Določili smo grafični način za prikaz rezultatov, s katerim čim bolje prikažemo čim več rezultatov hkrati.

Metoda vrne seznam rezultatov. Posamezne rezultate smo razdelili po skupinah odkritih vzorcev. Vsaki skupini je pripisan nabor mest, ki jih opiše. Za vsako skupimo smo izbrali do pet najboljših kombinacij prisotnosti proteinov, ki to skupino najbolje opisujejo. Slika 3.2 prikazuje kočni produkt vizualizacije. Na grafu je prikazana distribucija zaznanih interakcij kompleksa *SmB* na pozicijah od -40 do +40 nukleotidov relativno glede na mesta interakcije kompleksa *SmB*. Vsak vzorec vezave proteinov je prikazan z drugačno barvo. Vsi proteini znotraj

vzorca so povezani med seboj. Za vsak protein prikažemo tudi distribucijo zaznanih interakcij v okvirju velikosti 7 nukleotidov. Nad distribucijo smo v obliki kroga prikazali tudi skupno število interakcij, ki jih vzorec opiše.



Slika 3.2: Odkrite kombinacije proteinov, ki se pojavljajo v okolici mest vezave kompleksa *SmB*.

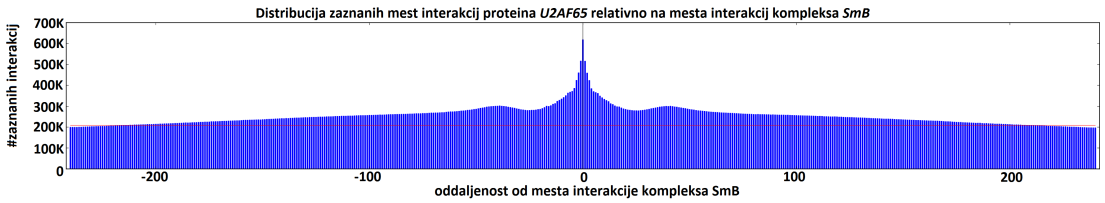
Poleg grafične predstavitve je narejena tudi predstavitev rezultatov v formatu CSV. S prikazom rezultatov v formatu CSV dobimo smiselno razporeditev stolpcev in se izognemo predolgemu poročilu o rezultatih. Prikaz v formatu CSV je razdeljen na dva dela. Prvi del predstavlja štiri datoteke, kjer vsaka opisuje eno genomsko področje. Drugi del je ena datoteka, kjer so vsa štiri področja združena skupaj. Razlika je v tem, da datoteka, ki združuje vse vzorce, ne more več prikazati posameznih skupin. S pomočjo te datoteke lahko opazujemo, kako so vzorci distribuirani glede na tip regije v primerjavi z ostalimi, in določimo, katero regijo najboljše opisujejo. V datoteke formata CSV smo vključili podatke o imenih proteinov znotraj vzorca, število pokritih pozicij tega vzorca in skupine, ki jim pripada, delež nukleotidov, ki jih vzorec opiše glede na nukleotide, ki jih opiše njegova skupina, in glede na vse opisane nukleotide. Genomski področji ekson-intron in intron-ekson imata še en dodaten stolpec, ki predstavlja dvojiški logaritem kvocienta deležev nukleotidov, ki jih opiše en vzorec, izmed vseh opisanih nukleotidov. Na ta način lažje primerjamo pripadnost enega vzorca posameznim regijam.

Poglavje 4

Rezultati in vrednotenje

4.1 Interakcije proteinov v okolici vezavnih mest kompleksa *SmB*

Za izbor kakovostne množice vhodnih podatkov smo uporabili vizualizacijo, s pomočjo katere smo določili vrednosti vrste parametrov. Slika 4.1 prikazuje distribucijo zaznanih mest interakcij proteina *U2AF65*, relativno na mesta interakcij kompleksa *SmB*, ki je prikazan v izhodišču grafa. Za izris grafa 4.1 smo določili več parametrov: območje, znotraj katerega opazujemo distribucijo ostalih proteinov, je 250 nukleotidov levo in desno od mesta interakcije kompleksa *SmB*. Znotraj tega območja smo izbrali področje, kjer distribucija najbolj odstopa od pričakovane enakomerne porazdelitve. Tako smo izbrali parameter **okvir**, znotraj katerega merimo gostoto vezavnih mest posameznih proteinov. Isti graf smo lahko uporabili tudi pri izbiri parametra **min_cDNA** (glej podpoglavje 3.1.1). Iz grafa na sliki 4.1 in grafov porazdelitev ostalih proteinov je bilo razvidno, da je najbolj informativno območje v okolici ±40 nukleotidov stran od mesta interakcije kompleksa *SmB* in RNA.



Slika 4.1: Distribucija zaznanih mest interakcij proteina *U2AF65* relativno na mesta interakcij kompleksa *SmB*.

Poleg grafov distribucije interakcij smo za izbiro parametra **min_cDNA** upora-

bili rezultate ponovitev poskusov. Vse parametre razen **min_cDNA** smo fiksirali in preverjali rezultate pri vrednostih **min_cDNA** od ena do tri. Tabela 4.1 prikazuje razmerje deležev opisanih mest, ki jih proteina *U2AF65* in *TIA* opisujeta v regijah ekson-intron in intron-ekson. Odnos je izračunan med pravilno in nasprotno regijo (razmerje p/n). Pri proteinu *TIA* smo delili delež opisanih pozicij v regiji ekson-intron z deležem v regiji intron-ekson. Pri proteinu *U2AF65* smo delili delež opisanih pozicij v regiji intron-ekson z deležem v regiji ekson-intron. Za ta proteina smo se odločili, ker sta edini znani komponenti kompleksa *SmB* in zanju točno vemo, kje in v kolikšni frekvenci ju lahko pričakujemo.

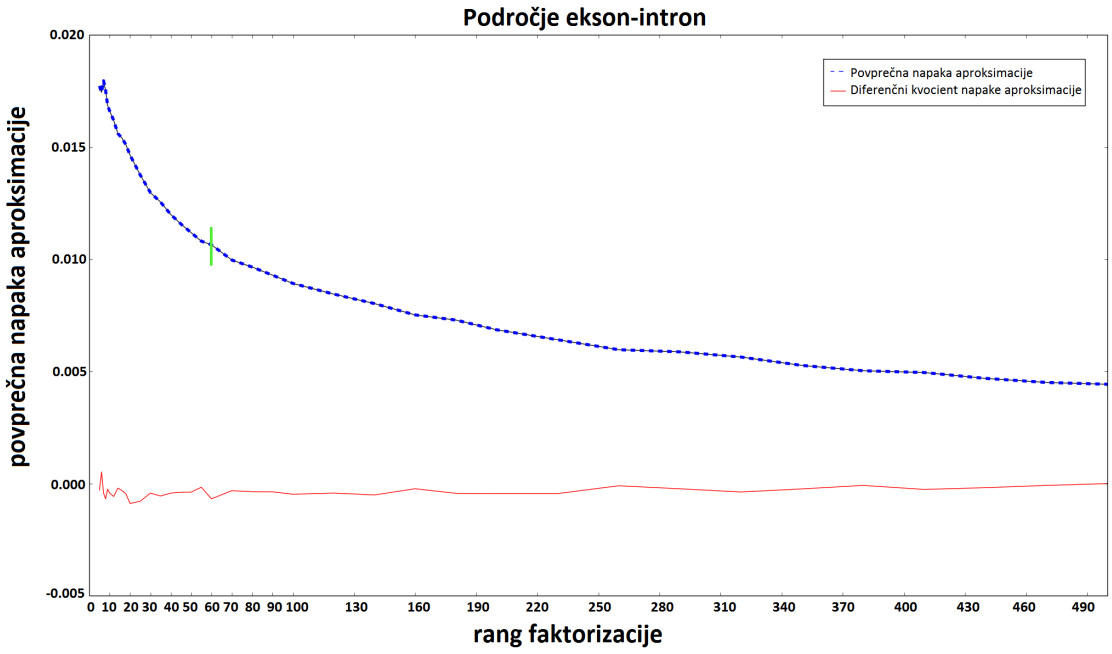
protein	p/n $min\_cDNA = 1$	p/n $min\_cDNA = 2$	p/n $min\_cDNA = 3$
U2AF65	13.26	6.95	0.58
***U2AF65* rnd**	NA	8.73	NA
TIA	1.62	2.08	6.05
***TIA* rnd**	NA	0.87	NA

Tabela 4.1: Razmerje deležev opisanih pozicij v pravilni in napačni genomski regiji za originalne in naključno permutirane podatke (" *U2AF65* rnd"in " *TIA* rnd") kompleksa *SmB*.

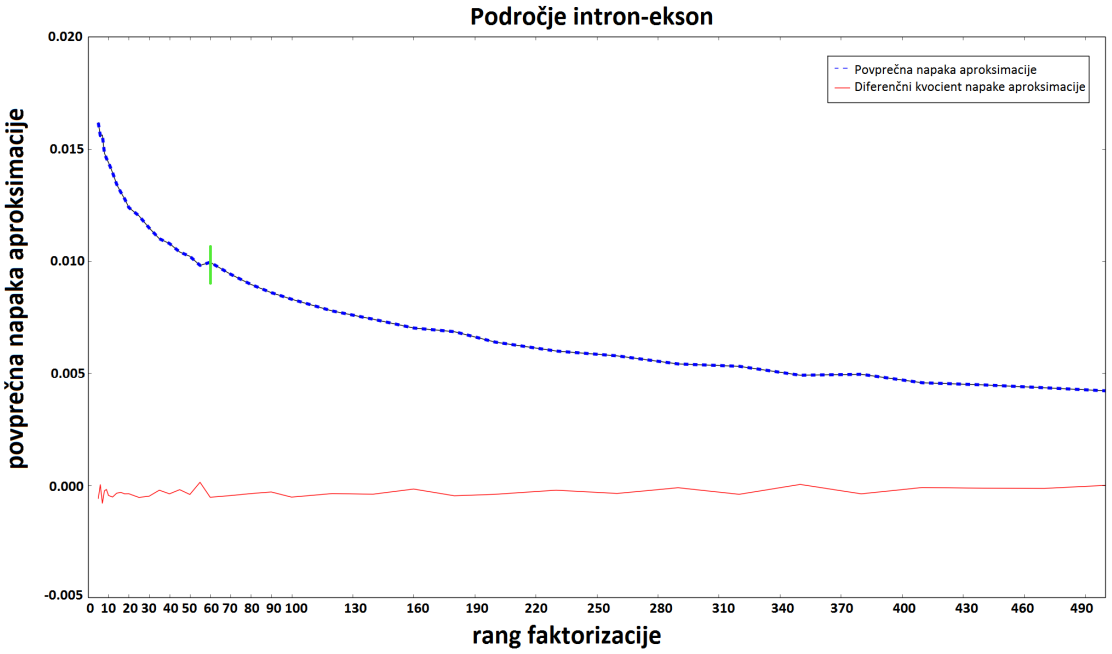
Iz tabele 4.1 je razvidno, da je na naključno permutiranih podatkih *SmB* pri proteinu *TIA* razmerje blizu vrednosti ena. Ta rezultat pomeni, da se protein *TIA* malenkost bolj pojavlja v regiji intron-ekson kot v regiji ekson-intron, kar pa ni pravilno oziroma ni pričakovani rezultat.

4.2 Izbira ranga faktorizacije

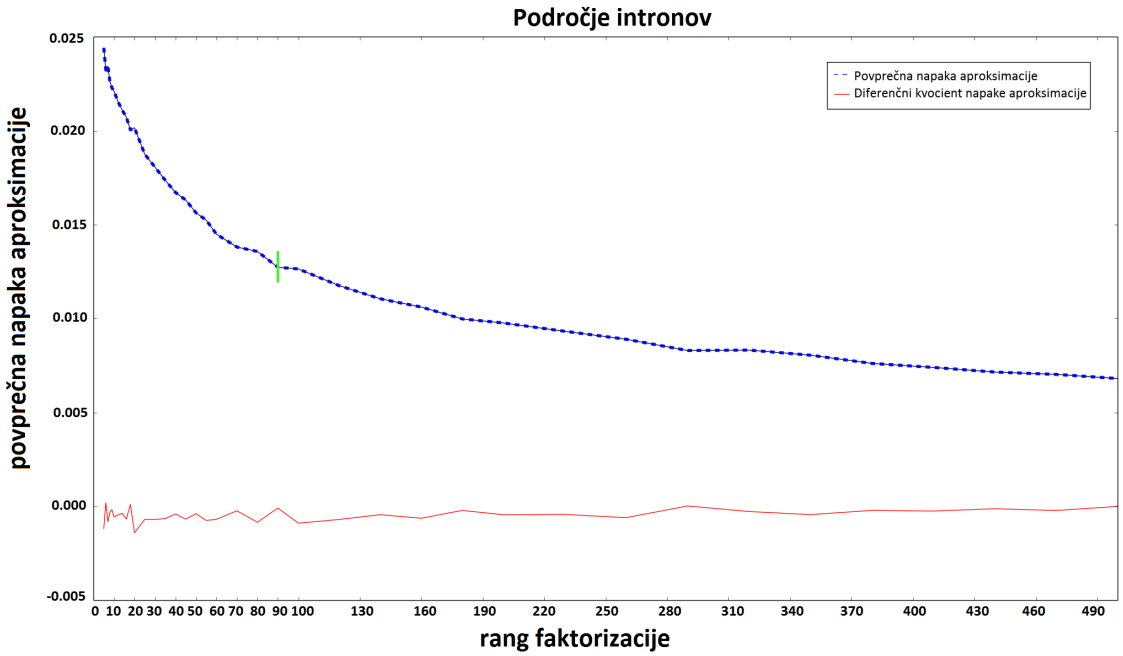
Pri nenadzorovanem učenju po navadi ne vemo dejanskega števila razredov. Zaradi tega je treba uporabiti postopek za iskanje najverjetnejšega števila razredov. Vsi postopki, ki podatke delijo v skupine, delujejo na principu zmanjševanja napake aproksimacije. Tudi pri NMF poskušamo dobiti najboljšo aproksimacijo prvotne matrike. Zaradi nekaterih lastnosti postopka NMF (glej podpoglavje 3.2) lahko NMF uporabimo kot orodje za iskanje skupin podobnih podatkov. Število pričakovanih skupin je treba podati vnaprej, in sicer z določitvijo ranga faktorizacije. Zaradi tega smo pri iskanju ranga morali preizkusiti nabor različnih vrednosti in izbrati najboljšega s pomočjo grafa napake aproksimacije (glej podpoglavje 3.2.3 in slike 4.2, 4.3, 4.4, 4.5).



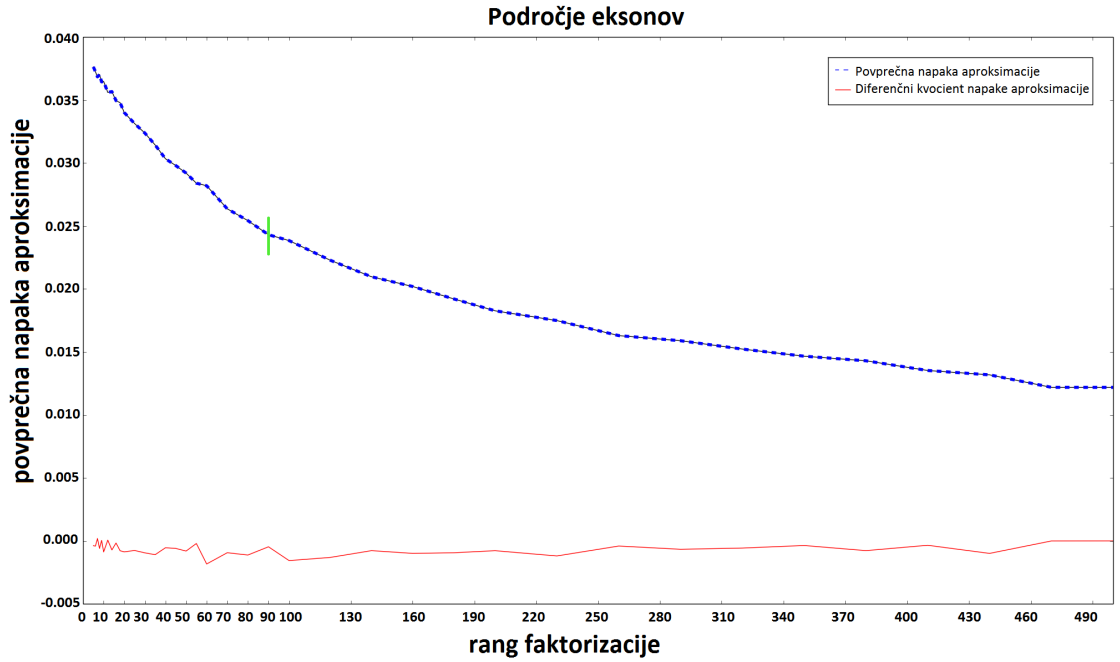
Slika 4.2: Povprečna napaka aproksimacije glede na rang faktorizacije v genomskem področju ekson-intron.



Slika 4.3: Povprečna napaka aproksimacije glede na rang faktorizacije v genomskem področju intron-ekson.



Slika 4.4: Povprečna napaka aproksimacije glede na rang faktorizacije v genomskem področju intronov.



Slika 4.5: Povprečna napaka aproksimacije glede na rang faktorizacije v genomskem področju eksonov.

S slik 4.2, 4.3, 4.4 in 4.5 je razvidno, da se hitrost padanja grafa pri regijah ekson-intron in intron-ekson spremeni v območju vrednosti ranga med 50 in 60. Ker v tem območju ni nobenih hitrih sprememb, smo se odločili izbrati za rang vrednost 55, ki je srednja vrednost območja. V področjih intronov in eksonov pa se hitro padanje grafa ustavi pri rangu 90, kar smo tudi izbrali kot rang faktorizacije teh območij.

4.3 Primerjava prisotnosti znanih proteinov kompleksa *SmB* v različnih genomskih področjih

Raziskava kompleksa *SmB* je izmed vseh proteinov, ki so del tega kompleksa, do sedaj odkrila samo dva proteina. Razlika v interakcijah proteinov *U2AF65* in *TIA* se najbolje opazi pri stiku eksonov in intronov. Protein *U2AF65* je znan po povečanem številu vezav na stiku med intronom in eksonom na traku 3' (angl. intron-exon junction). Protein *TIA*, za razliko od *U2AF65*, ima preferenco do področja, kjer se stikata ekson in intron na 5'-koncu (angl. exon-intron junction). To lastnost dveh proteinov smo uporabili kot kontrolo za našo metodo. Tabela 4.2 prikazuje rezultat naše metode za proteina *U2AF65* in *TIA*.

protein	% opisanih pozicij v ei	% opisanih pozicij v ie
U2AF65	0.44	3.11
***U2AF65* rnd**	0.31	2.12
TIA	3.31	1.59
***TIA* rnd**	1.24	1.08

Tabela 4.2: Delež opisanih pozicij v regiji ekson-intron in regiji intron-ekson za originalne ter naključno permutirane podatke (" *U2AF65* rnd"in " *TIA* rnd") kompleksa *SmB*.

Iz tabele 4.2 je razvidno, da *TIA* pokrije več kot dvakrat več pozicij na področju ekson-intron in *U2AF65* sedemkrat več pozicij na področju intron-ekson. Pri tem je protein *TIA* prisoten v 38 različnih kombinacijah, *U2AF65* pa v 90 različnih kombinacijah. Pri proteinu *U2AF65* so bile tri kombinacije prisotne na obeh področjih, pri proteinu *TIA* pa samo ena kombinacija.

Tabela 4.3 prikazuje štiri kombinacije proteinov *U2AF65* in *TIA*, ki smo jih odkrili v obeh genomskih regijah. Tri kombinacije proteina *U2AF65* jasno prikazujejo, da je ne glede na pojavitev v regiji ekson-intron še vedno večji delež opisanih

protein	log_2 ratio	% opisanih pozicij v ei	% opisanih pozicij v ie
***U2AF65* (-39 → -32)**	-0.080	0.15	0.15
***U2AF65* (-39 → -32) rnd**	NA	0.01	0.08
***U2AF65* (-32 → -25)**	-0.639	0.13	0.21
***U2AF65* (-32 → -25) rnd**	NA	0.01	0.07
***U2AF65* (-25 → -18)**	-3.955	0.03	0.47
***U2AF65* (-25 → -18) rnd**	NA	0.03	0.07
***TIA* (-18 → -11)**	-0.853	0.22	0.39
***TIA* (-18 → -11) rnd**	-0.540	0.09	0.12

Tabela 4.3: Dvojiški logaritem kvocienta deležev in deleža opisanih pozicij v regiji ekson-intron in regiji intron-ekson za proteine, ki se pojavljajo v obeh regijah. Prikaz rezultatov originalnih in naključno permutiranih podatkov (" *U2AF65* (-39 → -32) rnd, *TIA* (-18 → -11) rnd, itn.) kompleksa *SmB*.

interakcij v regijah intron-ekson. Pri proteinu *TIA* se samo ena kombinacija sopojavlja v obeh regijah in ima večji delež opisanih pozicij v regiji intron-ekson, ki pa še vedno ni znatno večji. Zraven imena vsakega proteina je v oklepajih napisano tudi področje nukleotidov, na katerem je ta kombinacija zaznana. Interakcije smo združevali v manjše skupine zaporednih nukleotidov zaradi zmanjševanja števila kombinacij. Vse štiri kombinacije imajo skupno lastnost, da so zaznane dokaj daleč od mesta interakcije kompleksa *SmB* z RNA, kar tudi zmanjšuje njihovo značilnost za to pozicijo.

V ostalih dveh regijah se oba proteina pojavljata z dosti večjim deležem. Tabela 4.4 prikazuje deleže opisanih interakcij v intronih in eksonih.

protein	% opisanih pozicij v gi	% opisanih pozicij v ge
U2AF65	25.03	39.83
***U2AF65* rnd**	36.57	54.75
TIA	23.42	26.11
***TIA* rnd**	4.68	9.91

Tabela 4.4: Delež opisanih pozicij v intronu in eksonu za originalne ter naključno permutirane podatke (" *U2AF65* rnd"in " *TIA* rnd") kompleksa *SmB*.

Področja intronov so zelo malo raziskana. Čeprav se nanje veže velika večina proteinov, je razumevanje njihove vloge zelo pomanjkljivo. Na eksone se prav tako veže mnogo proteinov, ki so del spajalnega telesca. To se odraža tudi v izjemno velikem številu opisanih pozicij, ki ga opisujeta oba proteina.

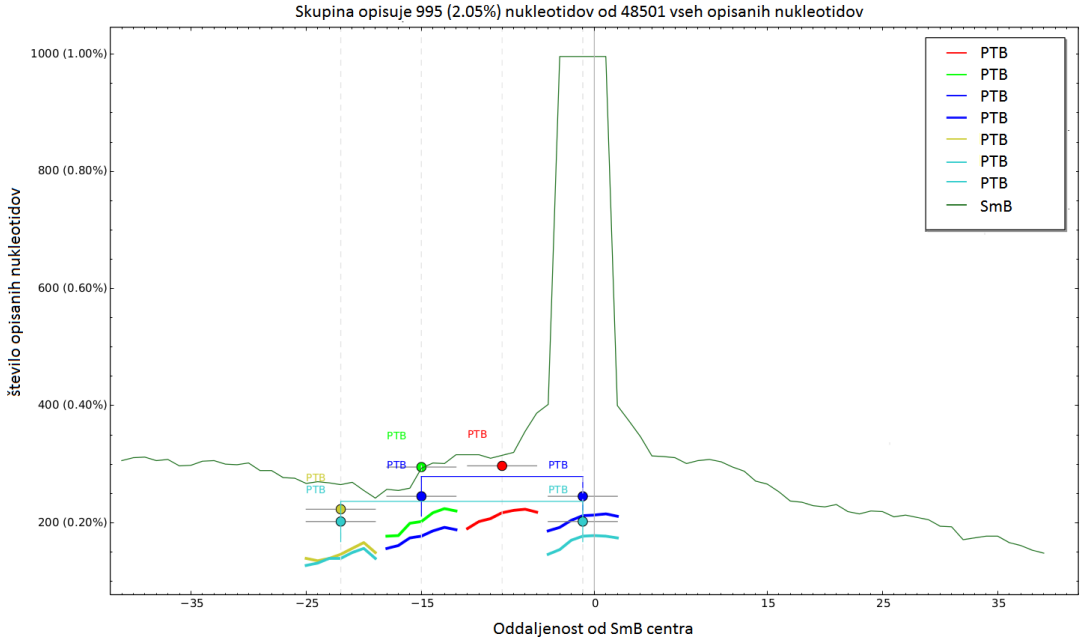
4.4 Vzorci interakcije v okolici mest vezave kompleksa *SmB*

Pri iskanju najboljših vzorcev smo skupine sortirali glede na število nukleotidov, ki jih opišejo. Skupine smo potem ročno pregledali in določili najbolj zanimive glede na proteine, ki se v njej vežejo, in postavitev vezav proteinov glede na distribucijo vezav kompleksa *SmB* (na grafih prikazano s temno zeleno barvo). Vzorce vsake genomske regije smo obravnavali posebej.

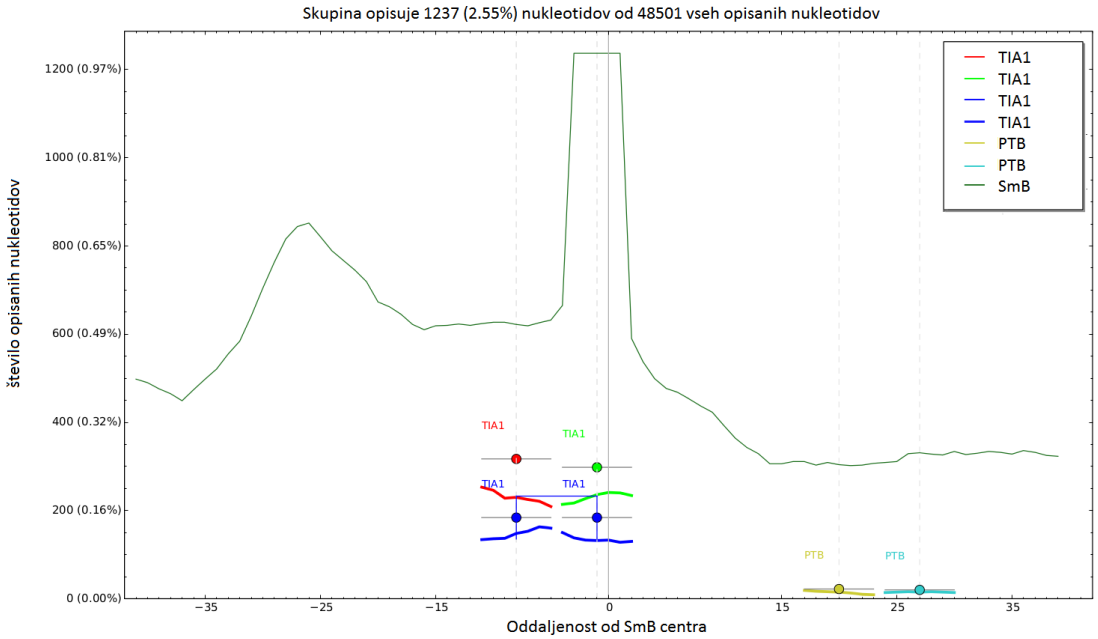
4.4.1 Področje ekson-intron

V regijah ekson-intron je prisoten protein *TIA* kot že znana komponenta kompleksa *SmB*. Pri iskanju zanimivih vzorcev smo iskali ujemanja z distribucijo kompleksa *SmB* in vzorce, v katerih se pojavlja protein *TIA*. Protein *PTB* se je pokazal kot zelo dober primer, ki ustreza obema kriterijema. Poleg v eni skupini, ki je bila sestavljena samo iz kombinacij proteina *PTB*, je bil *PTB* prisoten še v treh drugih skupinah skupaj s proteinom *TIA*. Poleg sopojavitve se je kot pomembna lastnost pokazala razdalja med zaznani interakcijami. Mesto interakcij proteina *PTB* je bilo vedno okrog 20 nukleotidov desno od mesta interakcije proteina *TIA*. Slike 4.6,4.7,4.8 in 4.9 prikazujejo štiri skupine, v katerih se pojavlja protein *PTB*.

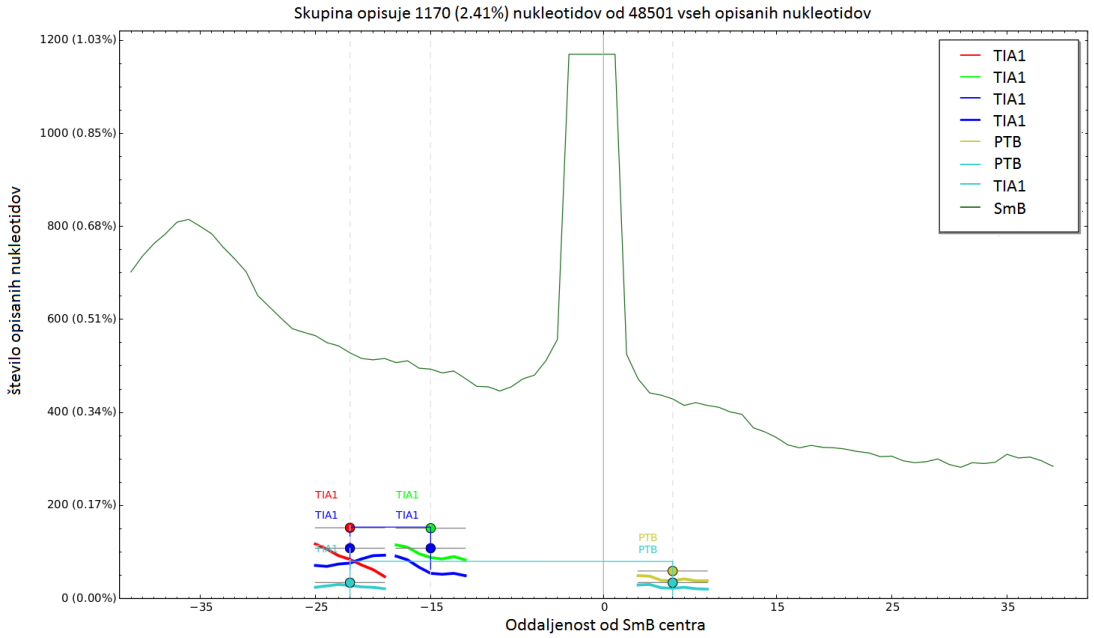
Protein *HuR* se je pokazal za zanimivega. Pojavlja se v štirih skupinah in opisuje veliko pozicij. Zanimiv je zaradi distribucije, ki se v vseh štirih skupinah popolnoma ujema z lokalnimi ekstremi distribucije kompleksa *SmB*. Protein *HuR* je znan po tem, da sodeluje pri regulaciji določenih komponent, ki vplivajo na formiranje in vezavo spajalnega telesca. Slike 4.10, 4.11, 4.12 in 4.13 prikazujejo štiri skupine, v katerih se pojavlja protein *HuR*.



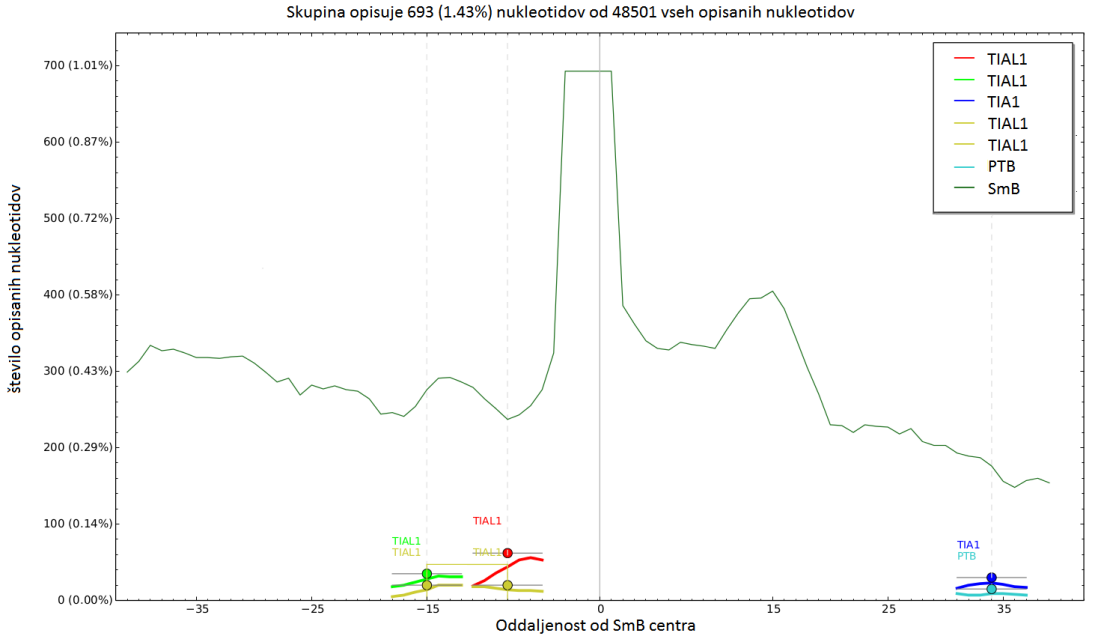
Slika 4.6: Odkrite kombinacije proteina *PTB*, ki se pojavljajo v okolici mest vezave kompleksa *SmB*.



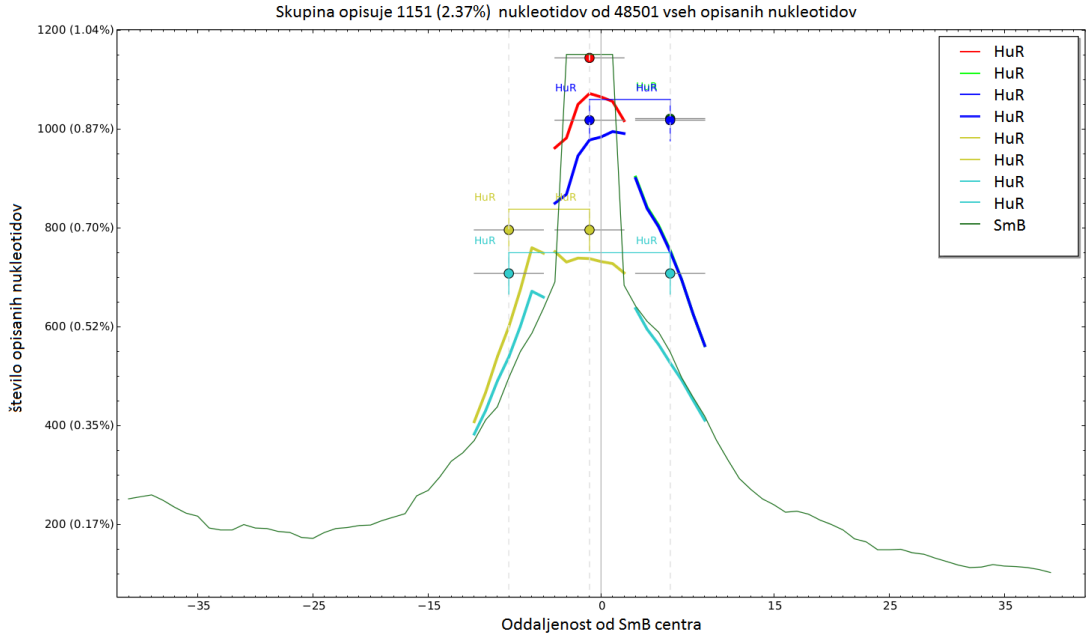
Slika 4.7: Odkrite kombinacije proteina *PTB*, ki se pojavljajo v okolici mest vezave kompleksa *SmB*.



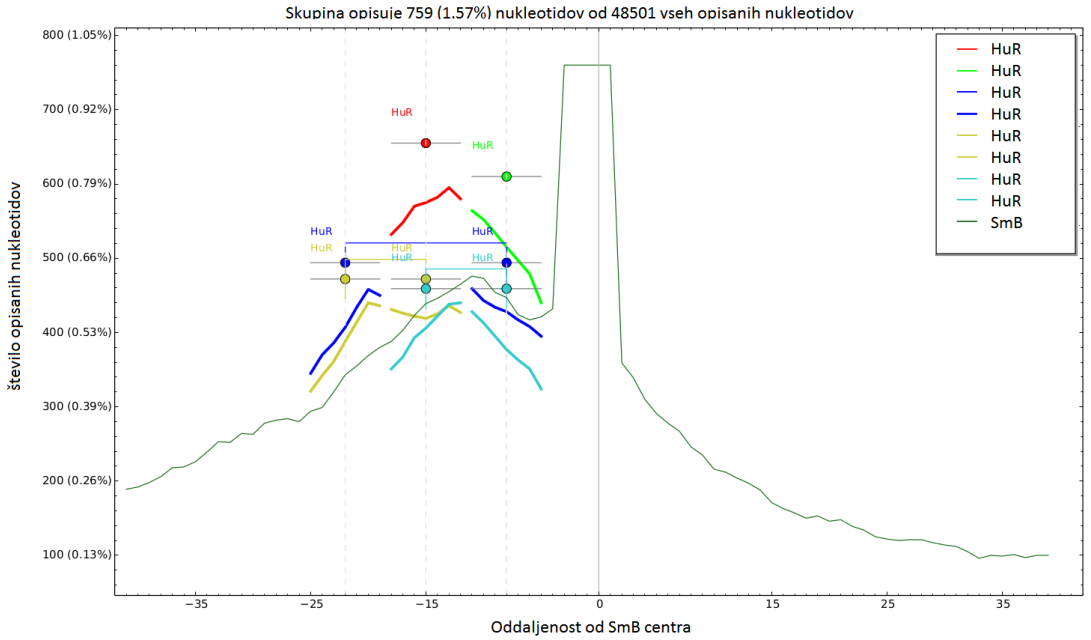
Slika 4.8: Odkrite kombinacije proteina *PTB*, ki se pojavljajo v okolici mest vezave kompleksa *SmB*.



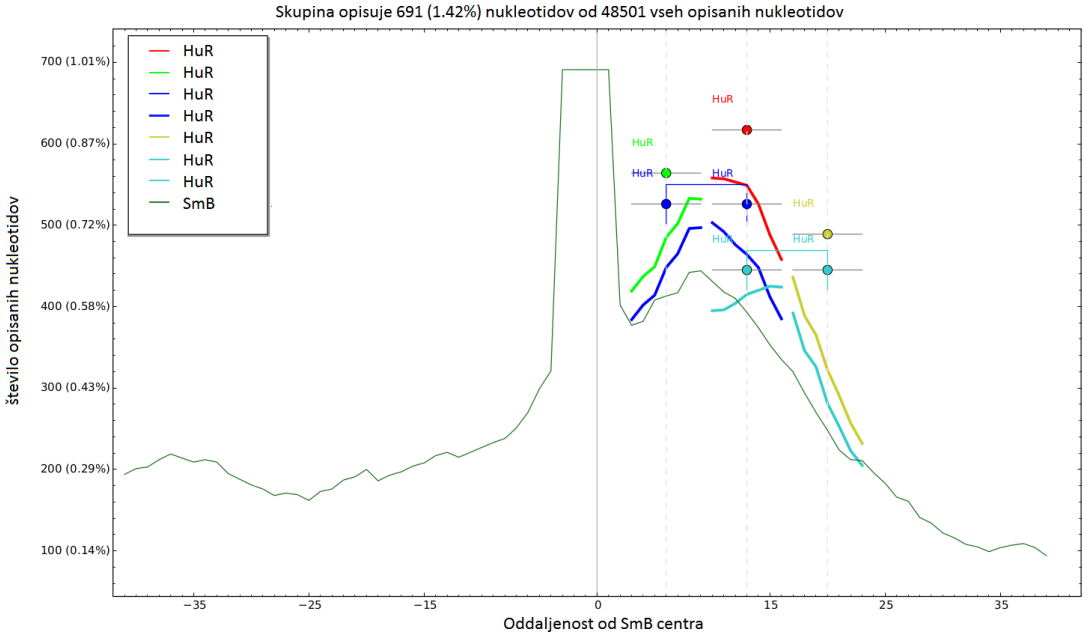
Slika 4.9: Odkrite kombinacije proteina PTB, ki se pojavljajo v okolici mest vezave kompleksa *SmB*.



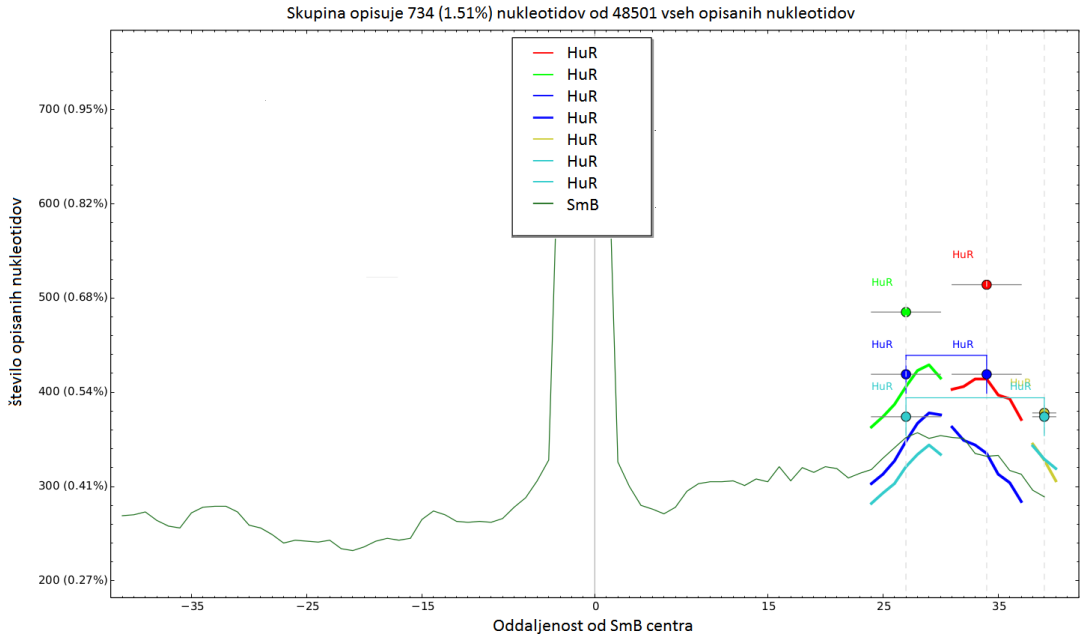
Slika 4.10: Odkrite kombinacije proteina *HUR*, ki se pojavljajo v okolici mest vezave kompleksa *SmB*.



Slika 4.11: Odkrite kombinacije proteina *HUR*, ki se pojavljajo v okolici mest vezave kompleksa *SmB*.



Slika 4.12: Odkrite kombinacije proteina *HUR*, ki se pojavljajo v okolici mest vezave kompleksa *SmB*.



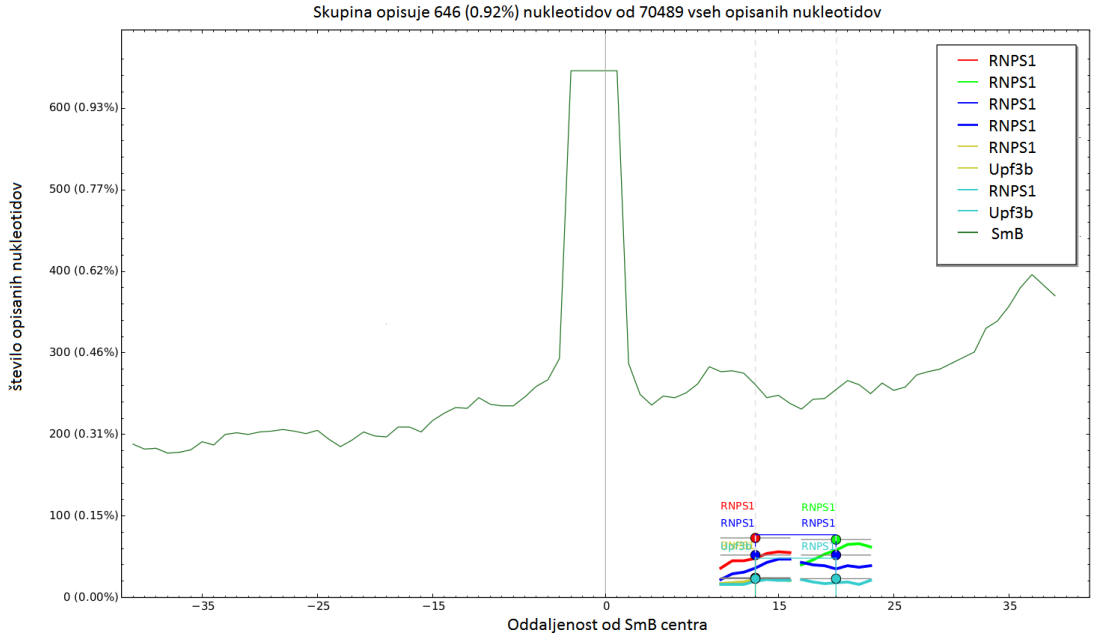
Slika 4.13: Odkrite kombinacije proteina *HUR*, ki se pojavljajo v okolici mest vezave kompleksa *SmB*.

4.4.2 Področje intron-ekson

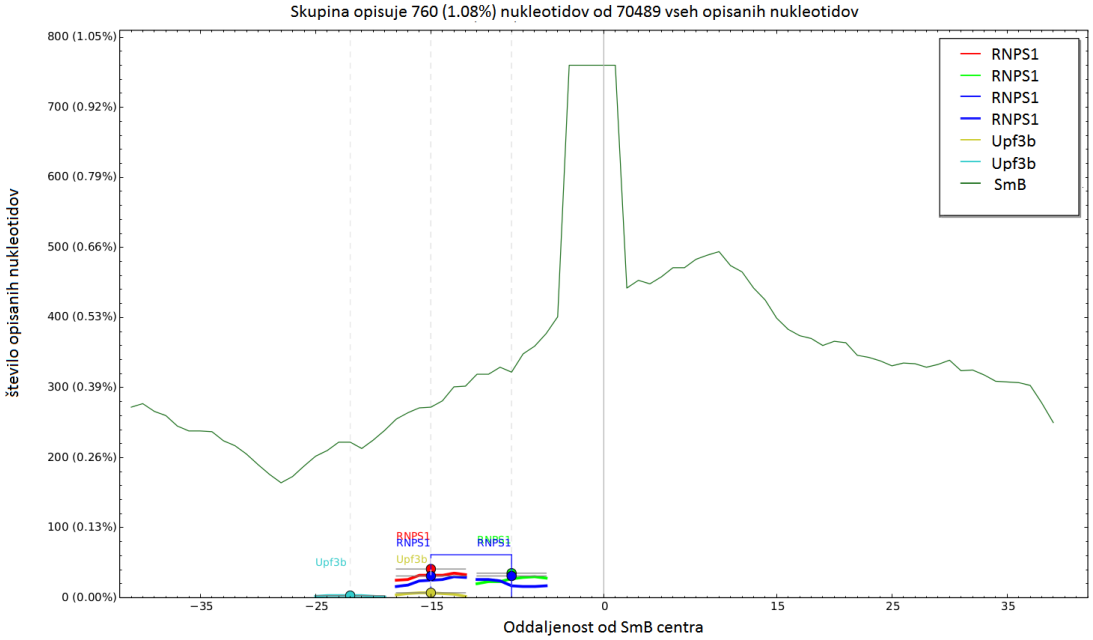
Na 3'-koncu so bili vzorci, ki vključujejo znane dele kompleksa *SmB*, manj izraženi. Protein *U2AF65* se je večinoma pojavljal v kombinacijah, kjer je bil edini prisoten. Edini protein, ki smo ga detektirali v kombinaciji s proteinom *U2AF65*, je bil *Musashi*. Ta se je vedno pojavljal na fiksni razdalji 10 nukleotidov levo od proteina *U2AF65*. Razen zanimive postavitve proteinov je bilo število opisanih pozicij premajhno, da bi ta vzorec določili za pomembnega.

Proteina, ki sta se izkazala kot bolj zanimiva, sta *RNPS1* in *Btz*. Pri kombinacijah, ki vključujejo protein *RNPS1*, smo opazili zanimiv pojav pri distribuciji interakcij kompleksa *SmB*. Ta je vedno dosegala lokalni ekstrem 20 nukleotidov desno od mesta interakcije proteina *RNPS*. Kot kažejo slike 4.14, 4.15, 4.16, 4.17 in 4.18, protein *RNPS1* opisuje samo del vseh nukleotidov, ki jih opisuje ta skupina. To lahko pomeni, da kompleks *SmB* vpliva na vezavo proteina *RNPS1*. Možna je tudi obratna razlaga, vendar bi to pomenilo, da je bila eksperimentalna detekcija proteina *RNPS1* slabše izvedena kot detekcija kompleksa *SmB*.

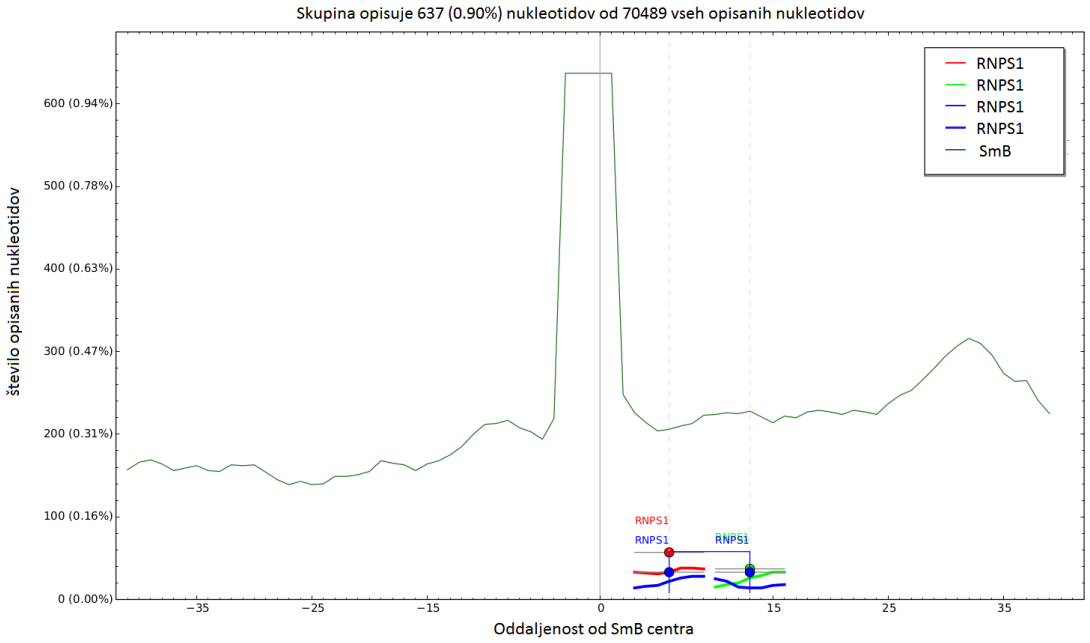
Protein *Btz* se v nekaterih primerih sopojavlja z zelo visokimi frekvencami (glej slike 4.19, 4.20, 4.21, 4.22). Delež interakcij, ki jih *Btz* v takšnih skupinah opiše, je nad 95%. Protein *Btz* je del kompleksa EJC (angl. *exon-junction complex*), ki sodeluje pri posttranskripcijskem uravnavanju mRNA.



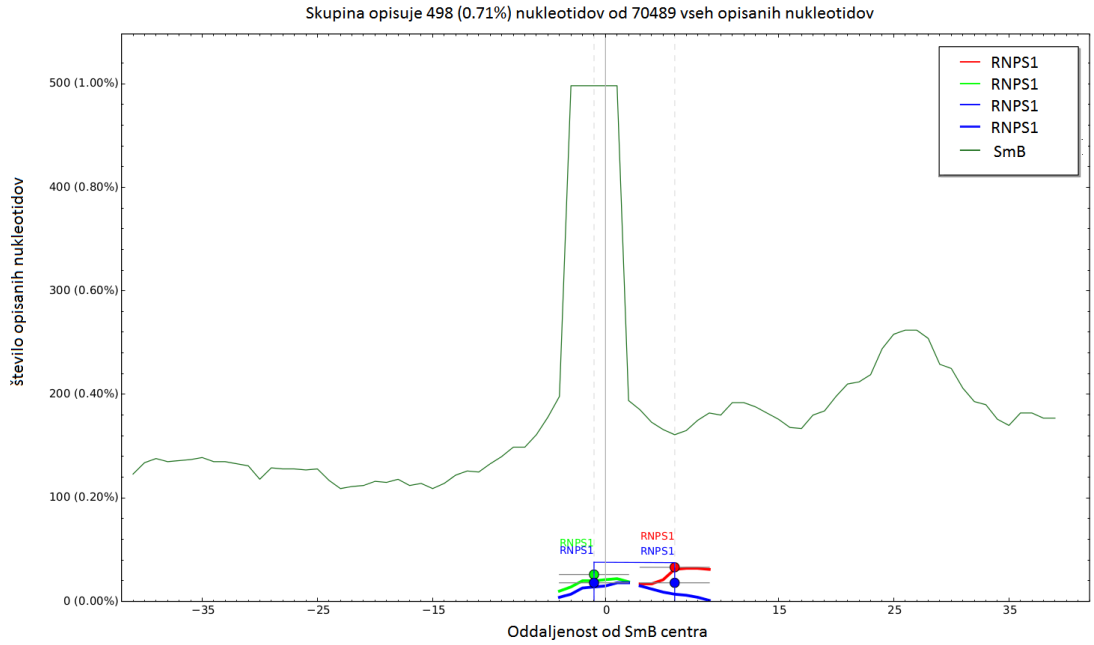
Slika 4.14: Odkrite kombinacije proteina *RNPS1*, ki se pojavljajo v okolici mest vezave kompleksa *SmB*.



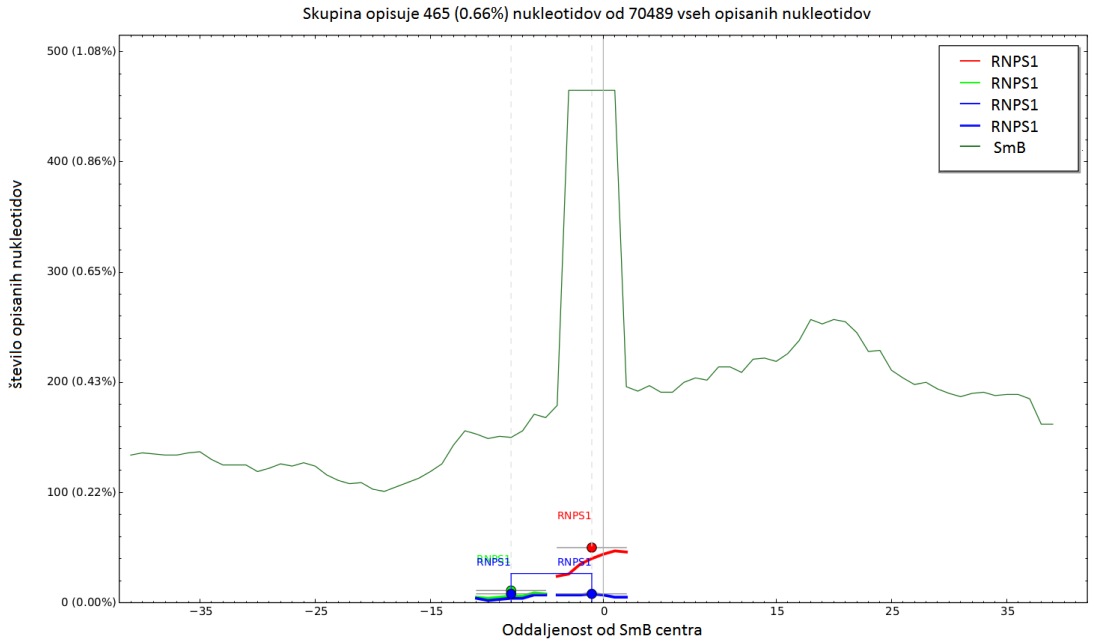
Slika 4.15: Odkrite kombinacije proteina *RNPS1*, ki se pojavljajo v okolici mest vezave kompleksa *SmB*.



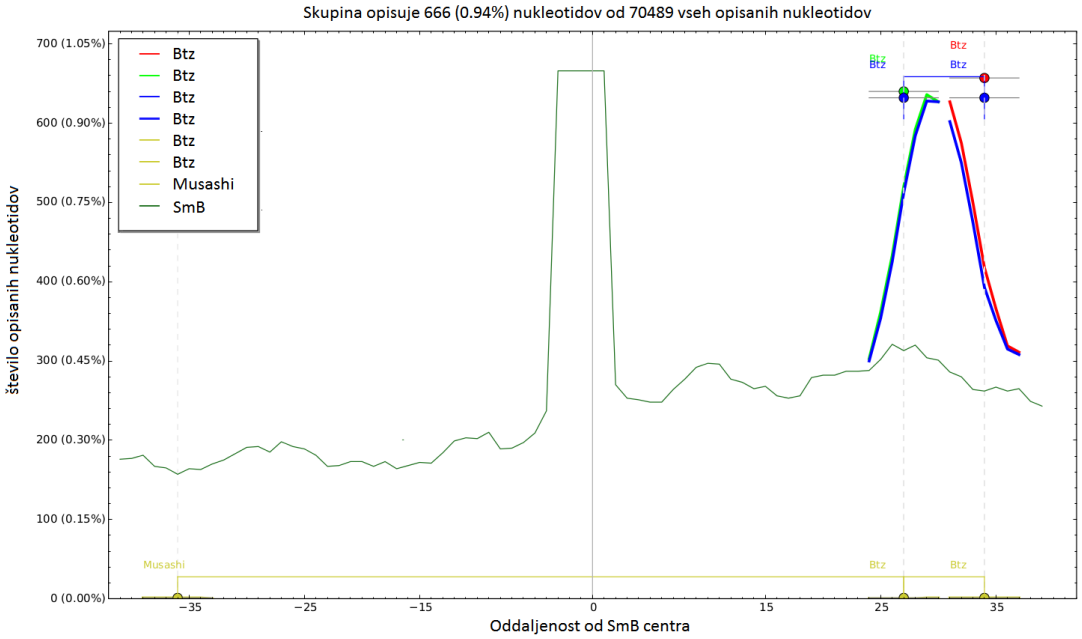
Slika 4.16: Odkrite kombinacije proteina *RNPS1*, ki se pojavljajo v okolici mest vezave kompleksa *SmB*.



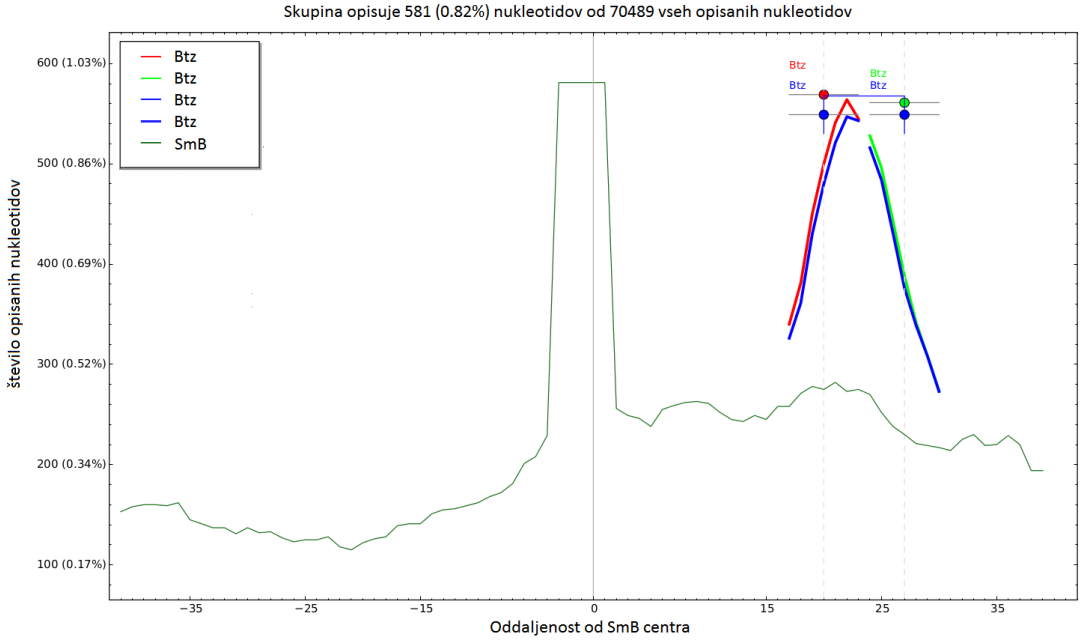
Slika 4.17: Odkrite kombinacije proteina *RNPS1*, ki se pojavljajo v okolici mest vezave kompleksa *SmB*.



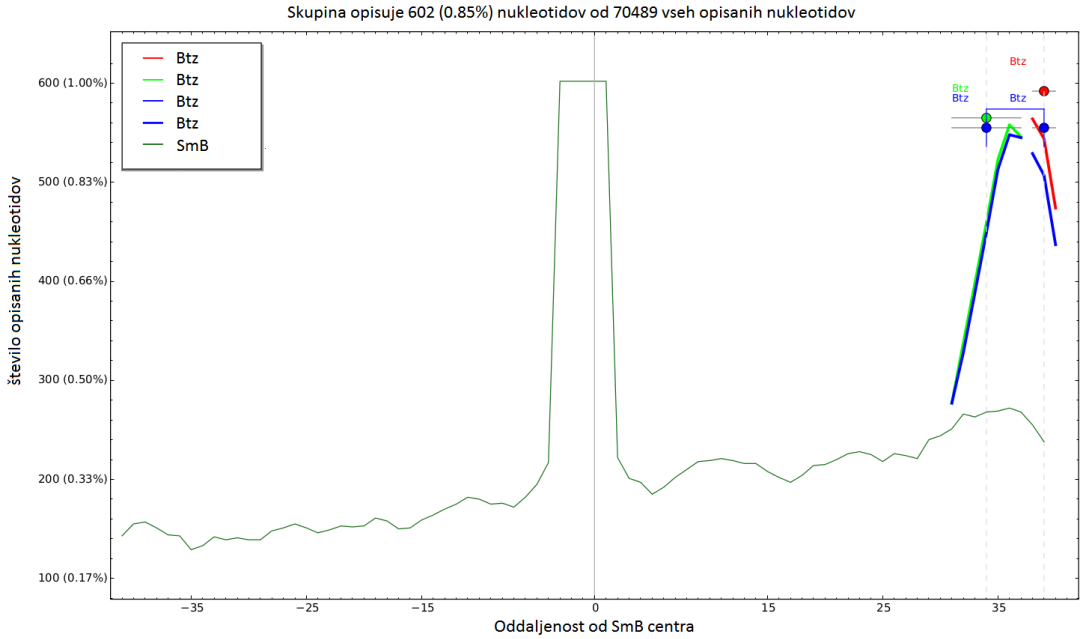
Slika 4.18: Odkrite kombinacije proteina *RNPS1*, ki se pojavljajo v okolici mest vezave kompleksa *SmB*.



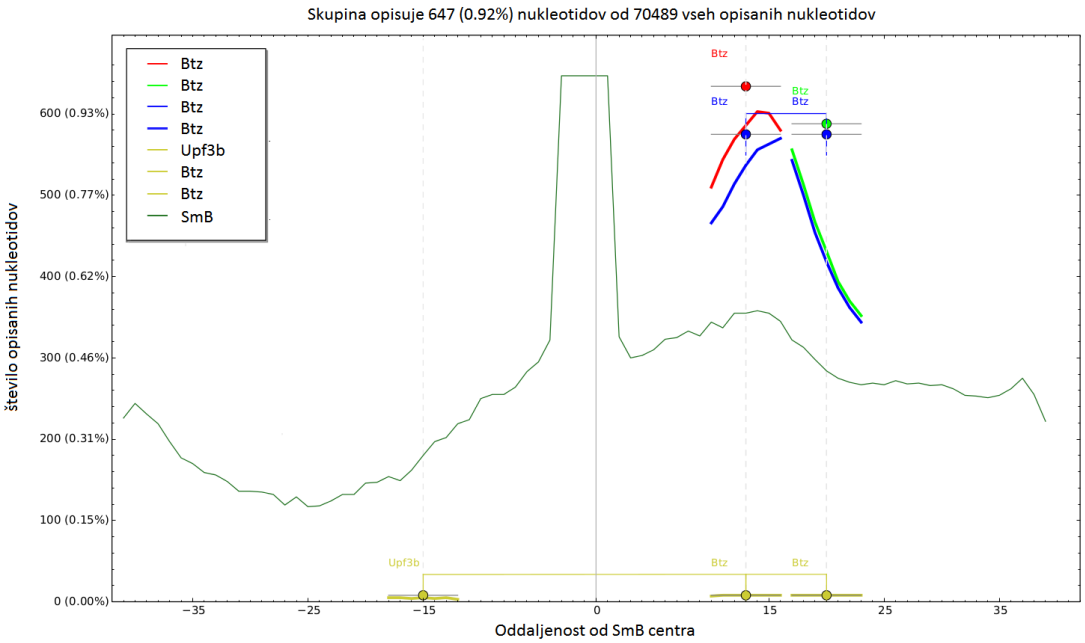
Slika 4.19: Odkrite kombinacije proteina *Btz*, ki se pojavljajo v okolici mest vezave kompleksa *SmB*.



Slika 4.20: Odkrite kombinacije proteina *Btz*, ki se pojavljajo v okolici mest vezave kompleksa *SmB*.



Slika 4.21: Odkrite kombinacije proteina *Btz*, ki se pojavljajo v okolici mest vezave kompleksa *SmB*.



Slika 4.22: Odkrite kombinacije proteina *Btz*, ki se pojavljajo v okolici mest vezave kompleksa *SmB*.

4.4.3 Področje intronov

Področje intronov je znano po tem, da je zelo popularno za veliko različnih proteinov, ampak je tudi najmanj raziskano. Skupine, formirane na tem področju, imajo lastnost, da so njihove kombinacije sestavljene večinoma samo iz enega proteina in da imajo zelo visok delež opisanih pozicij, glej tabelo 4.5. Poleg znanih komponent kompleksa *SmB* so najbolj pogosti proteini *hnRNPA1*, *HuR*, *PTB* in *hnRNPC*. Za te je znano, da sodelujejo v regulaciji na mestih mRNA, kamor se povezuje tudi spajalno telesce.

protein	% opisanih pozicij v gi	% opisanih pozicij v skupini
hnRNPA1	42.12	93.89
***hnRNPA1* rnd**	34.76	65.26
HuR	26.74	94.66
***HuR* rnd**	39.70	73.35
PTB	22.67	84.25
***PTB* rnd**	21.30	40.48
hnRNPC	21.09	91.61
***hnRNPC* rnd**	35.07	35.89

Tabela 4.5: Delež opisanih pozicij v skupini in regiji intronov za originalne in naključno permutirane podatke (*hnRNPA1* rnd, *HuR* rnd, itn.) kompleksa *SmB*.

Visok delež opisanih interakcij v skupini, ki ga imajo vsi proteini v tabeli 4.5, kaže na povezavo med kompleksom *SmB* in temi proteini.

4.4.4 Področje eksonov

Na področju eksonov se kot na področju intronov veže veliko različnih proteinov. Zaradi te lastnosti nam ne predstavlja zanimivega področja, ker je zelo težko določiti pomembne vzorce. Večina vseh odkritih vzorcev se pojavlja z zelo visoko frekvenco. Proteini, ki se najbolj pogosto pojavljajo, so enaki kot pri področju intronov (glej tabelo 4.5).

Poglavje 5

Sklepne ugotovitve

Na začetku diplomske naloge smo si zadali dva cilja. Razvoj metode za detekcijo komponent proteinskih kompleksov v interakciji z RNA smo uspešno dokončali. Začetni nabor vhodnih podatkov smo uspešno obdelali in zmanjšali na najbolj informativno podmnožico. Pri tem postopku smo uporabili več različnih parametrov, ki smo jih morali odkriti. S faktorizacijo nenegativnih matrik smo razdelili podatke v optimalno število skupin in za vsako poiskali vse vzorce mest vezave proteinov v okolici vezave kompleksa *SmB*. Na koncu smo razvili metodo za prikazovanje rezultatov, nato pa smo preverili še delovanje metode na realnih podatkih o kompleksu *SmB*.

Za doseganje vseh ciljev smo uporabili predznanje iz molekularne biologije in strojnega učenja. Diplomsko delo je zahtevalo veliko različnih poskusov in konzultacij z eksperti biologi. Odprlo je veliko novih vprašanj in prikazalo izjemno kompleksnost problema razumevanja interakcij proteinskih kompleksov. Omogočilo nam je, da ugotovimo napake v podatkih in postopkih, ter pustilo možnost za nadaljnjo raziskavo.

5.1 Pomembnost kakovosti podatkov

Uporabljeni podatki so bili različnih kakovosti. Razlika v kakovosti je bila še najbolj opazna, če smo primerjali podatke, pridobljene z različnimi metodami in v različnih laboratorijih. Podatki, ki smo jih uporabili, so bili pridobljeni v enem laboratoriju z istim postopkom, ampak v različnih poskusih. V vsakem poskusu so bili uporabljeni drugačni parametri, pa tudi obseg sekvenciranja podatkov je bil različen. Z vzorčenjem podatkov na enako število enot smo poskusili izničiti pristranskosti, ki jih potencialno lahko vnašajo eksperimentalne razlike.

V nekaterih primerih se je zgodilo, da je metoda določene logične skupine

razbila na več ločenih skupin. K temu je vodila predvsem slaba kakovost podatkov. Majhna učinkovitost povezovanja s svetlobo UV v metodi iCLIP je imela velik vpliv na odkrivanje skupin (glej poglavje 2.3). Pri vsaki metodi je zelo pomembno, da so podatki čim bolj kakovostni, vendar boljše metode lahko iz manj kakovostnih podatkov več odkrijejo. Naša metoda uporablja samo informacijo o mestu interakcije, kar nikakor ni dovolj, da bi odkrili logične skupine, ker so podatki razpršeni na več RNA-jih. V nadaljnjem razvoju metode bi bilo smiselno omogočiti vključevanje dodatnih informacij, s katerimi bi omilili vzroke za zgoraj omenjene težave.

5.2 Uporabnost predznanja

Pri razvoju in vrednotenju algoritmov smo uporabili veliko predznanja iz strojnega učenja, matematike in molekularne biologije. Pri predprocesiranju podatkov poskusov smo uporabili biološko predznanje. Iz podatkov smo izločili proteine, za katere je znano, da ne vstopajo v kompleks *SmB*, in se tako izognili nepotrebnemu računanju. Pri določanju vrednosti parametrov smo prav tako uporabili predznanje o bioloških lastnostih proteinov in tudi nasvete avtorjev metode iCLIP. Pri izbiri parametrov je bilo treba razumeti tudi način delovanja matrične faktorizacije in vpliv spremembe različnih parametrov na rezultate, ki jih vrne postopek faktorizacije. Pri vrednotenju rezultatov smo ponovno uporabili predznanje iz biologije ter tako preverili, ali odkrite kombinacije vsebujejo pričakovane skupine proteinov v okolici kompleksa *SmB*.

5.3 Iskanje komponent proteinskih kompleksov

Na rezultate iskanja komponent proteinskih kompleksov vpliva mnogo parametrov. Odkriti vzorci so bili večinoma sestavljeni iz enega ali dveh različnih proteinov. Ni nam uspelo odkriti nove komponente kompleksa *SmB*. Vseeno smo dobili zanimive rezultate, ki smo jih lahko preverili, in na ta način ovrednotili predlagano metodo. Veliko najdenih kombinacij je vsebovalo že znane proteine, ki so del drugih kompleksov in vplivajo na vezavo kompleksa *SmB*. Poleg informacije o sestavi vzorca je bila pomembna tudi informacija o deležu opisanih pozicij. Pri komponentah kompleksa *SmB* smo poročali še o razmerju med deleži opisanih pozicij ($log_2 ratio$) (glej podpoglavje 3.5). Pri vseh meritvah smo poročali o rezultatih permutacijskega testa, s čimer so rezultati postali bolj jasni. Permutacijski test je pokazal, da se odkriti vzorci značilno razlikujejo od rezultatov na naključnih podatkih. To

nakazuje pravilnost delovanja metode in pravilnost njenih rezultatov.

5.4 Nadaljnje delo

Diplomsko delo je odprlo veliko novih vprašanj. Kompleksnost problema je daleč presegla okvire diplomske naloge. Hkrati je ustvarila veliko načrtov za nadaljnje delo. Slabi podatki so nam pokazali, da bo treba uporabiti več različnih virov in izkoristiti vse možnosti, ki jih ponuja NMF. Zanimivo bi bilo preučiti, kako se detektirana mesta interakcij spreminjajo skozi različne stopnje fragmentiranja. Metodo bi bilo vredno preveriti tudi na podatkih o drugih proteinskih kompleksih (npr., kompleks proteinov EJC).

Literatura

[1] König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B. & Ule, J. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nature structural & molecular biology, 17(7), 909-915.

[2] König, J., Zarnack, K., Luscombe, N. M. & Ule, J. (2012). Protein–RNA interactions: new genomic technologies and perspectives. Nature Reviews Genetics, 13(2), 77-83.

[3] Žitnik, M (2012) A Matrix Factorization Approach for Inference of Prediction Models from Heterogeneous Data Sources. EngD thesis, Univerza v Ljubljani).

[4] Kambach, C., Walke, S., Young, R., Avis, J. M., de la Fortelle, E., Raker, V. A. & Nagai, K. (1999). Crystal structures of two Sm protein complexes and their implications for the assembly of the spliceosomal snRNPs. Cell, 96(3), 375-387.

[5] I. Kononenko. Strojno ucenje. Zalozba FE in FRI, 2005.

[6] Clark, D. P. & Russell, L. D. (2010). Molecular biology made simple and fun. Cache River Press.

[7] Wikipedia (2014). Non-negative matrix factorization, dostopno na: http://en.wikipedia.org/wiki/Non-negative_matrix_factorization

[8] Protein Data Bank (PDB), dostopno na: http://www.rcsb.org/pdb/explore/explore.do?structureId=2Y9C/

[9] Briese, M., Sibley, C., Haberman, N., Wang, Z., König, J., Perera, D., Wickramasinghe, V.O., Venkitaraman, A.R., Smith, C.W.J., Curk, T., Ule, J. (2014). Genome-wide analysis of spliceosomal interactions identifies distinct classes of branch points, poslano v recenzijo.

[10] BedGraph Track Format, UCSC, dostopno na:
https://genome.ucsc.edu/goldenPath/help/bedgraph.html